Multiple Hypothesis Testing and Causal Discovery

PhD Thesis

Phillip Bredahl Mogensen
August 2022

Department of Mathematical Sciences,
University of Copenhagen.

This thesis has been submitted to the PhD School of the Faculty of Science,
University of Copenhagen

Phillip Bredahl Mogensen
pbm@math.ku.dk
phmo1406@gmail.com

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark

Academic advisors	Associate Professor Bo Markussen (primary) University of Copenhagen
	Professor Helle Sørensen University of Copenhagen
	Associate Professor Shyam Gopalakrishnan University of Copenhagen
Assessment committee	Professor Niels Richard Hansen (chair) University of Copenhagen
	Professor Ingeborg Waernbaum University of Uppsala
	Professor Jelle Goeman University of Leiden
Submission date	August 7, 2022
ISBN	978-87-7125-060-2

Preface

This thesis contains a collection of the scientific work I have carried out as a PhD student under the supervision of Bo Markussen and Helle Sørensen (Department of Mathematical Sciences), and Shyam Gopalakrishnan (Globe Institute), all from the University of Copenhagen.

Nine years ago, I had just finished gymnasium and found myself with no ambitions of pursuing any academic studies. On a whim, I enrolled at the University of Copenhagen anyway – just to try it out. As it turns out, that impulsive decision is, to date, the best I have ever made. The University of Copenhagen is where I discovered my love of mathematics, made friends whom I cherish deeply, met excellent colleagues and, finally, where I met my fiancé.

To my friends and family, who have far too often been neglected over the stress of my studies, I am grateful that you continue to care. Your friendship and support has meant, and continues to mean, the world to me.

To my coworkers: Lasse, Martin, and Nikolaj, thank you for enduring a worldwide pandemic with me and bringing laughs, even while we were stuck in our apartments and it was difficult to find joy in our studies; Alex, Christian, Debbie, Jonas, Leonard, Matthieu, Nena, Nicola, Niklas, Sebastian, Shimeng, Predrag, and many others, thank you for making our offices a joyous place to be. To the senior faculty, I would like to thank Niels Richard Hansen, Anders Tolver, and Jonas Peters for always taking the time to give me academic advice – or just the time to enjoy a cup of coffee together. The Department of Mathematical Sciences at UCPH enjoys a special culture in which PhDs, postdocs and professors are equal. This, I have an enormous appreciation for.

I thank my supervisors, Helle Sørensen and Shyam Gopalakrishnan, who have guided me academically and provided valuable ideas at times when I had no idea how to continue my research. I especially thank my primary supervisor, Bo Markussen, for continually inspiring me to explore on my own, pointing me in the right direction when I got lost, and always supporting me.

Finally, I am exceptionally grateful to Anna, whose love, kindness, and generosity is such that I can not repay it in a single lifetime. Without it, I would have given up long before ever reaching the finish line.

Phillip Bredahl Mogensen
August 2022

Abstract

This PhD thesis deals with two different subjects: multiple hypothesis testing and causal discovery.

In the first part of the thesis, we propose a new family of combination tests – called the 'Too Many, Too Improbable' (TMTI) test – for combining evidence from multiple hypothesis tests into a single test of a joint hypothesis. We then prove that the proposed family of tests fits within a larger family of tests, for which we prove a quadratic shortcut for carrying out Closed Testing Procedures. Finally, we show empirically that a subfamily of the proposed family can be easily approximated, facilitating the use of these tests in large-scale studies.

In the second part of this thesis, we consider the task of learning causal graphs from data. First, we attempt to learn finite summary graphs of infinite-dimensional graphs of discrete-time stochastic processes. We develop simple algorithms that score the existence of causal links by aggregating local linear effects and validate these algorithms on data from a case competition. However, we argue that the observed high performance of these algorithms may be inflated by the presence of an artifact in simulated data. Next, we propose a novel method – called the Invariant Ancestry Search – for learning causal ancestors of a response variable using data sampled from heterogeneous environments. We prove that the proposed method recovers subsets of ancestors of the response with high probability, if given infinite amounts of data, and we show empirically that the guarantees hold approximately when applied to finite samples.

Resumé

Denne PhD afhandling berører to forskellige emner: multipel hypotese testing og kausal læring.

I første del af afhandlingen foreslår vi en ny familie af kombinationstests – kaldet 'Too Many, Too Improbable' – til at kombinere evidens fra flere hypotese tests i et enkelt test af en simultan hypotese. Vi beviser at den foreslåede familie af tests ligger i en større familie af tests, for hvilken vi beviser, at en lukket testprocedure kan udføres med en kvadratisk genvej. Vi viser empirisk at en delfamilie af den foreslåede familie kan approksimeres nemt, hvilket muliggør brugen af disse tests i store studier.

I anden del af afhandlingen berører vi hvordan kausale grafer kan læres fra data. Først diskuterer vi, hvordan endelige resumégrafer af uendeligt store grafer for diskret-tids stokastiske processer kan læres fra data. Vi udvikler simple algoritmer, der scorer eksistensen af kausale sammenhænge ved at aggregere tilstedeværelsen af lokale lineære effekter, og vi validerer disse algoritmer på data fra en case competition. Vi argumenterer dog for, at den observerede høje ydeevne af disse er kunstigt hævet af en artefakt i simuleret data. Dernæst foreslår vi en ny metode – kaldet Invariant Ancestry Search – til at lære kausale forfædre til en responsvariabel fra data observeret på tværs af heterogene miljøer. Vi beviser at den foreslåede metode kan finde en delmængde af forfædre med høj sandsynlighed, hvis vi bliver givet uendelige mængder data, og vi viser empirisk at garantierne holder approksimativt når anvendt på endelige mængder data.

Contents

Chapter 1

Introduction

Causal discovery and multiple testing are two branches of mathematical statistics that are not directly related but not disjoint either. Causal discovery deals with the task of learning causal relations from data. These causal relations are often represented by Directed Acyclic Graphs (DAGs) (see, e.g., Lauritzen, 1996; Pearl, 2009; Peters et al., 2017). Multiple testing deals with how different types of errors, often the family-wise error rate (FWER) or false discovery rate (FDR) (see, e.g., Tukey, 1953; Benjamini and Hochberg, 1995), can be controlled when using data to make decisions about multiple hypotheses simultaneously. In learning causal structures, one often needs to test a hypothesis to determine the existence and direction of cause-effect relationships. If one intends to learn more than a single cause-effect relationship, i.e., if one wishes to learn (parts of) a DAG, one needs to test multiple hypotheses.[1] Thus, multiple testing deals with a general problem, which occurs in, among other places, causal discovery.

This thesis deals with these two subjects; multiple testing and causal discovery. In the present chapter, we briefly introduce both subjects before diving into them more deeply in Chapters 2 to 4.

1.1 Contributions and structure

We reference papers that are a part of this thesis in boldface capital letters, and we refer to the contents of a paper using the format 'paper reference, internal reference'. For example, we refer to Section 5 of Paper **A** by 'Paper **A**, Section 5'. The papers included in this thesis are:

Paper **A** Phillip B. Mogensen and Bo Markussen. 'Too Many, Too Improbable: testing joint hypotheses and closed testing shortcuts' arXiv preprint arXiv:2108.04731 (2021).

[1] There are exceptions to this. For example, Zheng et al. (2018) frame the problem of structure learning as a continuous optimization problem. However, some objections have been made towards this approach (Reisach et al., 2021; Seng et al., 2022).

Paper **B** Sebastian Weichwald, Martin E. Jakobsen, Phillip B. Mogensen, Lasse Petersen, Nikolaj Thams, Gherardo Varando. 'Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values' Proceedings of the NeurIPS 2019 Competition and Demonstration Track, PMLR 123:27-36, 2020.

Paper **C** Phillip B. Mogensen, Nikolaj Thams, Jonas Peters. 'Invariant Ancestry Search' Proceedings of the 39th International Conference on Machine Learning, PMLR 162:15832-15857, 2022.

Each paper is contained within its own chapter. In each chapter, we also discuss additional aspects that we either developed after submission or aspects that did not fit within the papers themselves. We include Paper **A** in the format in which it is available online, and we include Paper **B** and Paper **C** in the formats of the journals in which they have been accepted. Each chapter can be read independently of one another. Each paper contains a list of references made within that paper. A bibliography of references made outside each paper is available at the back of the thesis.

In Chapter 2, we present Paper **A**, in which we propose a new family of *combination tests* for testing joint hypotheses and prove that this family satisfies a shortcut for *closed testing*. Paper **A** has been submitted to the Journal of Statistical Planning and Inference and is, at the time of writing, still under review. In Chapter 3, we present Paper **B**. In this paper, we discuss how to learn *summary graphs* of time-homogeneous discrete-time stochastic processes. In addition, we discuss the effects of a particular simulation artifact of additive noise models when learning causal graphs. Paper **B** was an invited paper at the thirty-third Conference on Neural Information Processing Systems (NeurIPS) as the result of the team of authors winning a case competition at said conference. Lastly, in Chapter 4, we present Paper **C**. Here, we develop a novel method – called Invariant Ancestry Search – for learning *causal ancestors* of a response variable from data sampled from heterogeneous environments. This work was accepted into the thirty-ninth International Conference on Machine Learning (ICML).

1.2 Multiple testing

1.2.1 A motivating example

The following example is a simplified and slightly modified telling of a series of studies conducted by Anna L. Colliander while she was a PhD student at the Technical University of Denmark, Department of Health Technologies.

A researcher is developing novel compounds for cancer immunotherapy. The starting point for these novel compounds is the drug resiquimod, which

has been shown to be an effective anti-cancer drug due to its immunostimulating properties (Wu et al., 2004). However, resiquimod has never reached the clinic due to its severe systemic toxicity (Pockros et al., 2007). In an attempt to reduce the toxicity of resiquimod but retain its anti-cancer effect, the researcher develops six different resiquimod analogs (i.e., chemically modified versions of resiquimod). First, the researcher performs a study to investigate the anti-cancer properties of the drugs. This study shows, for each drug, non-inferiority with respect to resiquimod. Thus, informally, each drug works at least as well as resiquimod with respect to its anti-cancer properties. Next, the researcher wants to investigate whether the systemic toxicity is reduced.

To do so, the researcher sets up a mouse study: each analog is administered to eight tumor-bearing mice, and a blood sample is taken from each mouse two hours later. To measure the degree of systemic toxicity, the serum level of Interleukin-6 (IL-6) in each blood sample is recorded. For each group of mice, the researcher then compares the serum level of IL-6 to the levels in a positive control group of mice treated with resiquimod, using a one-sided T-test. The null hypothesis here is that the systemic toxicity of the respective drug is at least as high as the toxicity of resiquimod. This is tested against the alternative hypothesis that the analog is better. The p-values for these tests are shown in Table 1.1.

Drug	A	B	C	D	E	F
p-value	2.5%	4.9%	5.9%	6.7%	8.1%	42.5%

Table 1.1: Unadjusted p-values for the tests described in Section 1.2.1.

Upon seeing that none of the p-values are significant at the usual 5% level after Bonferroni correcting, the researcher concludes that none of the drugs worked as intended and decides not to move forward with the developed analogs. However, that conclusion is incorrect; the researcher was simply asking the wrong question – or rather, not enough questions. Instead, the researcher should have asked:

1. Did any of the drugs work?

2. How many of the drugs worked?

3. Which of the drugs worked?

In the first part of this thesis, we deal with methods of answering this 'ladder of questions' by using the framework of closed testing, introduced by Marcus et al. (1976), and the work on dissonant closed testing of Goeman and Solari (2011). In many cases, it turns out, these questions get more and more difficult to answer as we move down the ladder. That is, the more specific the question, the more stringent our demands for quality data must be.

Looking back at the p-values in Table 1.1, it seems unlikely that not a single drug had the intended effect. Even though we failed to reject anything after Bonferroni correcting, it seems unlikely that not a single drug had lower toxicity: there are simply *too many* of the p-values that are *too small* for all hypotheses to be true. Indeed, applying the framework of closed testing along with the test we develop in Paper **A**, we find that a p-value for the first question is 0.01%. Thus, there is compelling evidence that at least one of the drugs had lower systemic toxicity – but we do not know which one. Moving a step down the ladder, we can say (with 95% confidence) that at least four of the six drugs are significantly less toxic – but again, we can not say which ones. We can further refine this answer and say (with 95% confidence) that at least four of the drugs A through E are less toxic than resiquimod. Thus, we can eliminate drug F and conclude that most of the remaining drugs had the intended effect of reducing the systemic toxicity.

In this case, answering just the third question (with a Bonferroni correction) meant that potentially valuable research was thrown out. The lack of a significant finding was likely due to the study being underpowered, rather than none of the drugs having reduced systemic toxicity.

1.2.2 Hypotheses, tests and p-values

Throughout this section, let $(\Omega, \mathcal{F})$ be a measurable space and let $\mathcal{P}$ be a family of probability measures on $(\Omega, \mathcal{F})$. Suppose there exists a random variable X, which has distribution $\mathbb{P}_X \in \mathcal{P}$. The theory of hypothesis testing deals with the following problem: given an observation x of X, how do we determine whether to reject or accept a hypothesis about $\mathbb{P}_X$, and how do we quantify the certainty in our decision? In this section, we review some basic theory of hypothesis testing and introduce the problem of multiple testing.

We begin with the formal definition of a hypothesis.

Definition 1.1. *A hypothesis $H_0 \subseteq \mathcal{P}$ is a non-empty subset of distributions in $\mathcal{P}$. The interpretation of the hypothesis is that the true, data-generating distribution $\mathbb{P}_X$ lies in H_0. The alternative hypothesis H_A to H_0 is the complement of H_0 in $\mathcal{P}$, i.e., $H_A = \mathcal{P} \setminus H_0$.*

For every hypothesis, we desire a means to test it. That is, we wish to use data to determine whether or not the hypothesis H_0 should be rejected.

Definition 1.2. *A test of the hypothesis H_0 is a subset $R \subseteq \Omega$ of possible observations for which the hypothesis H_0 is rejected. The set R is called the rejection region. The complementary set $R^{\complement}$ is the set of observations for which we fail to reject the hypothesis.*

We can equivalently define a test by a decision rule $\phi : \Omega \to \{0, 1\}$ for H_0, given by

$$\phi(x) = \begin{cases} 0, & x \in R^{\complement} \\ 1, & x \in R \end{cases}.$$

When constructing a test of H_0, we are particularly interested in the size and power of the test.

Definition 1.3. *Let $\phi : \Omega \to \{0, 1\}$ be a test of H_0. The size of the ϕ is*

$$\alpha(\phi) \coloneqq \sup_{\mathbb{P} \in H_0} \mathbb{E}_{\mathbb{P}}\phi = \sup_{\mathbb{P} \in H_0} \mathbb{P}(H_0 \text{ is rejected}).$$

That is, the size of ϕ is the largest possible probability of falsely rejecting H_0 when it is true.

The power of ϕ under a given alternative measure $\mathbb{P}_A \in H_A$ is

$$\beta_\phi(\mathbb{P}_A) \coloneqq \mathbb{E}_{\mathbb{P}_A}\phi = \mathbb{P}_A(H_0 \text{ is rejected}).$$

That is, the power of a test is the probability of correctly rejecting H_0 when it is false.

Usually, a test of H_0 is constructed not only for a fixed size α but for all possible sizes $\alpha \in (0, 1)$. In this case, we can define the p-value for the test of H_0.

Definition 1.4. *Let $\phi_\alpha : \Omega \to \{0, 1\}$ be a family of decision rules for the test of H_0, each at size α, and let $x \in \Omega$ be a fixed observation. The p-value (if it exists) for the test of H_0 is given by*

$$p \coloneqq \inf\{\alpha \in (0, 1) \mid \phi_\alpha(x) = 1\}.$$

That is, the p-value for the test of H_0 is the smallest possible size α such that we reject the hypothesis H_0.

Lastly, we define an admissible test, following the definition of Lehmann (1947).

Definition 1.5. *We say that ϕ_α is an admissible test if, for a fixed $\alpha \in (0, 1)$, there exists no other test with the same size that has power uniformly greater than (or equal to) that of ϕ_α, but not identically equal to.*

In Example 1.1, we go through an example of how to construct the well-known two-tailed test for a population mean in the case with known variance.

Example 1.1 (Z-test with known variance)**.** *Let $X \sim N(\beta, \sigma^2)$ be a random variable with unknown mean β and known variance σ^2, and let $(X_i)_{i=1}^n$ be independent and identically distributed (i.i.d.) copies of X. Denote by $x =$*

$(x_i)_{i=1}^n$ an observation of $(X_i)_{i=1}^n$. The hypothesis of interest is that the mean of X is equal to some value $\mu_0 \in \mathbb{R}$. Here, the family of probability measures is $\mathcal{P} := \{N(\mu, \sigma^2) \mid \mu \in \mathbb{R}\}$, and the hypothesis $H_0 = \{N(\mu_0, \sigma^2)\}$. The alternative hypothesis is $H_A = \{N(\mu, \sigma^2) \mid \mu \in \mathbb{R} \setminus \{\mu_0\}\}$. We call H_0 a simple hypothesis because it consists only of a single measure.

Now, fix $\alpha \in (0, 1)$ and denote by $\bar{x}$ the empirical mean of $x_1, \ldots, x_n$. To construct a test with size α, we define the rejection region

$$R_\alpha = \left\{ x \in \mathbb{R}^n \mid \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \geq z_{1-\alpha/2} \right\},$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a $N(0, 1)$ distribution. Noting that the empirical mean of $(X_i)_{i=1}^n$ has distribution $N(\mu_0, \sigma^2/n)$ under the null (meaning that $(\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ has distribution $N(0, 1)$), it is trivial that the test with rejection region R_α has size α.

To get the p-value for this test, we find the rejection region with the smallest size p such that the observation x lies in R_p. We can find this p-value by setting

$$\left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| = z_{1-p/2}.$$

Solving for p and letting Φ denote the CDF of a standard Gaussian distribution, we find that

$$p = 2 \times \left(1 - \Phi \left(\left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \right) \right)$$

is a p-value for the test of H_0.

1.2.3 Error types and error control for a single test

Usually, when reporting the results of a hypothesis test, we report the p-value. We choose, a priori, a significance level α (often $\alpha = 0.05$), and we then compute a p-value for the hypothesis of interest. If the p-value falls below α, we reject H_0, and if the p-value is above α, we fail to reject H_0. Doing so has the benefit of controlling the probability of making a Type I error.

Definition 1.6. *Let H_0 be a hypothesis and H_A be its alternative. A Type I error is made if we reject H_0 when it is true. A Type II error is made if we fail to reject H_0 when it is false.*

We say that a p-value (or test) is valid at level α if it controls the Type I error at level α. If a p-value controls the Type I error at any α, we simply say that it is valid. That is, a valid p-value can equivalently be defined as a random variable P satisfying

$$H_0 \text{ true} \implies \forall \alpha \in (0, 1) : \mathbb{P}(P \leq \alpha) \leq \alpha. \tag{1.1}$$

Trivially, a valid p-value can be constructed by choosing $P \sim U(0,1)$. However, this is not a good choice of p-value because we have a high probability of making a Type II error when using it. Thus, a *good* p-value is one that is both valid and has high power – i.e., a low probability of making a Type II error. Informally speaking, maximizing power while controlling the Type I error simply means that we make the right choice as often as possible.

1.2.4 Error types and error control for multiple tests

So far, we have only discussed the testing of a single hypothesis. In this thesis, we are mainly concerned with scenarios where there are multiple hypotheses for which we must make decisions simultaneously.

Going forward, let $H_1, \ldots, H_m$ be $m \in \mathbb{N}$ hypotheses and let $P_1, \ldots, P_m$ be random variables such that each P_i satisfies Type I error control for the hypothesis H_i.

The basic problem of multiple testing is that even though the Type I error is controlled for each marginal hypothesis, there will be an excess of Type I errors when testing multiple hypotheses.[2] For example, if $P_1, \ldots, P_m$ are mutually independent and all m hypotheses are true, the probability of making at least one Type I error is

$$\mathbb{P}\left(\bigcup_{i=1}^{m}(P_i \leq \alpha)\right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^{m}(P_i > \alpha)\right) \geq 1 - (1-\alpha)^m > \alpha.$$

Thus, the probability of making at least one Type I error is larger than α and even goes to one in the limit of m. Hence, one cannot approach multiple testing in the same manner as when testing a single hypothesis. Instead, one typically adjusts the marginal p-values in some way to control a different error type. We define below some of the most common targets that one attempts to control.

Definition 1.7. *Let $H_1, \ldots, H_m$ be hypotheses. Let $T \subseteq \{1, \ldots, m\} =: [m]$ and $F := [m] \setminus T$ be the indices of the hypotheses that are true and false, respectively. Let $R \subseteq [m]$ be the indices of the hypotheses that are rejected by some procedure.*

The Familywise Error Rate (FWER) (Tukey, 1953) is the probability of falsely rejecting at least one true hypothesis:

$$\text{FWER} := \mathbb{P}(|T \cap R| \geq 1).$$

We say that the FWER is strongly controlled at level $\alpha \in (0,1)$ if $\text{FWER} \leq \alpha$ *for all possible constellations of T and F, and weakly controlled at level α if* $\text{FWER} \leq \alpha$ *only when $T = [m]$ (i.e., all marginal hypotheses are true).*

[2] Except in the (somewhat artificial) case where all hypotheses are perfectly dependent.

The k-FWER (Hommel and Hoffmann, 1988; Korn et al., 2004; Lehmann and Romano, 2005), or generalized FWER, is the probability of falsely rejecting at least k true hypotheses:

$$\text{k-FWER} \coloneqq \mathbb{P}(|T \cap R| \geq k).$$

The False Discovery Proportion (FDP) is the ratio of falsely rejected true hypotheses to the number of rejected hypotheses (set to zero when no rejections are made),

$$\text{FDP} \coloneqq \begin{cases} \frac{|T \cap R|}{|R|}, & |R| > 0 \\ 0, & |R| = 0 \end{cases}.$$

The False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) is the expected FDP,

$$\text{FDR} \coloneqq \mathbb{E}(\text{FDP}).$$

Above, the expectations and probabilities are with respect to the data-generating process of the underlying data.

The literature on methods to control the targets in Definition 1.7 is vast and still expanding. For instance, the FWER can be controlled by employing a Bonferroni correction (Dunn, 1961), which is equivalent to testing each marginal hypothesis at level α/m. Hommel and Hoffmann (1988); Korn et al. (2004); Lehmann and Romano (2005) develop methods for controlling the k-FWER, and Benjamini and Hochberg (1995) develop a simple method for controlling the FDR. Lehmann and Romano (2005) also develop methods that control the tail probability of the FDP, i.e., they develop methods such that $\mathbb{P}(\text{FDP} \geq \gamma) \leq \alpha$ for any choice of $\gamma, \alpha \in (0, 1)$. Note that controlling the k-FWER for $k \in \mathbb{N}$ for a rejection set R is equivalent to controlling the tail-probability of the FDP at $\gamma = k/|R|$. Thus, to control one is to control the other. In this thesis, we generally discuss the k-FWER, but the reader should note that any statement we give on k-FWER control can be equivalently formulated as a statement on controlling the tail-probability of the FDP.

1.2.5 Closed Testing Procedures

Closed Testing Procedures (CTPs), introduced by Marcus et al. (1976), have garnered much attention since their introduction in 1976. We formally define CTPs in Paper **A**, Section 5. Briefly, a CTP rejects a hypothesis $H_{\mathcal{J}} \coloneqq \bigcap_{j \in \mathcal{J}} H_j$ if and only if every joint hypothesis that contains $H_{\mathcal{J}}$ is rejected. Such procedures are generally accepted to be more powerful than their non-closed counterparts. This was formalized by Sonnemann and Finner (1988), who showed that every admissible test that controls the FWER is a closed procedure. Indeed, most well-known procedures controlling the FWER can be shown to be closed (e.g., Bonferroni corrections, Holm step-down corrections, Šidák corrections, etc.). Romano et al. (2011) further showed that only

consonant procedures (see Definition 2.1) are admissible for FWER control. However, Goeman and Solari (2011) argue that the FWER is, in many cases, too strict a target to control. The authors instead use *dissonant*[3] procedures (formally defined in Definition 2.1) to control the k-FWER and compute $1-\alpha$ confidence sets for the number of false hypotheses in any given rejection set. Later, Goeman et al. (2021) showed that only CTPs are admissible for controlling the tail FDP (and thereby k-FWER). That is, any procedure that is not closed but controls the k-FWER is either equivalent to a closed procedure or dominated by a closed procedure. Thus, in many cases, the optimal choice of multiple testing procedure is a closed one. A particularly attractive property of CTPs, as shown in Goeman and Solari (2011), is that they allow for valid *post hoc* inference. That is, the k at which one wishes to control the k-FWER can be chosen after reviewing the data without compromising the level at which the k-FWER is controlled.

Despite the many attractive properties of CTPs, these can be difficult to compute due to their exponential complexity in the number of marginal tests. For m marginal tests, one must perform $2^m - 1$ tests to complete a CTP. This is not unreasonable for small m, but even at $m = 30$, the number of hypotheses to test exceeds a billion. Therefore, researchers looking to perform a CTP attempt to find *shortcuts* in their procedures – that is, a way to reduce the number of tests to something more manageable. As it turns out, this is often possible (Romano et al., 2011). Many of these shortcuts are identical in construction: when testing all joint hypotheses of size m' that contain a hypothesis $H_{\mathcal{J}}$, it suffices to consider only the p-values in $\mathcal{J}$ combined with the largest of the remaining p-values. For instance, Zaykin et al. (2002) provide this shortcut for their test, the Truncated Product Method (TPM). Dobriban (2020) provides the same shortcut for a family of tests that are monotone and symmetric, and Tian et al. (2021) give a similar shortcut for test statistics that are sums of marginal tests. Goeman et al. (2019) provide a shortcut for determining lower bounds on the proportion of true discoveries in a rejection set S in $\mathcal{O}(m \log m)$ time for Simes type local tests. Additionally, the authors show that bounds for all subsequent rejection sets S' can be found in $\mathcal{O}(|S'|)$ time. Goeman et al. (2021) extends this shortcut to the Higher Criticism test of Donoho and Jin (2004).

In Paper **A**, we give CTP shortcuts for a family of combination tests that are monotone, but not necessarily symmetric. This shortcut allows, for example, the user to compute adjusted p-values for all marginal hypotheses in quadratic time. Additionally, it can be used to compute $1 - \alpha$ confidence sets for the number of false hypotheses in a rejection set S in $\mathcal{O}(m|S|)$ time when $|S| < m$ and $\mathcal{O}(m)$ time when $|S| = m$. In Section 2.8.2, we argue that by using a binary search, the number of marginal hypotheses that can be rejected with FWER control can be identified in $\mathcal{O}(m \log m)$ time. Similarly,

[3]Also called non-consonant procedures in the literature.

when employing a binary search, $1-\alpha$ confidence sets for the number of false hypotheses in S can be identified in $\mathcal{O}(m \log |S|)$ time when $|S| < m$ and $\mathcal{O}(\log m)$ time when $|S| = m$. Lastly, using a binary search allows us to identify the largest possible rejection set that controls the k-FWER at a fixed k in $\mathcal{O}(m(\log m)^2)$ time.

1.2.6 Joint hypothesis testing and combination testing

In closed testing, it is necessary to test joint hypotheses, i.e., the hypothesis $H_{\mathcal{J}}$ that all marginal hypotheses in $\mathcal{J}$ are simultaneously true. Constructing a test of $H_{\mathcal{J}}$ is no different than constructing a test for any marginal hypothesis: we must construct a rule which takes data and rejects $H_{\mathcal{J}}$ with probability at most α, given that $H_{\mathcal{J}}$ is true. In addition, we aim to construct this rule such that $H_{\mathcal{J}}$ is rejected as often as possible when $H_{\mathcal{J}}$ is not true.

The construction of a joint test can be done in many ways. For instance, in multiple linear regression, hypotheses about multiple coefficients can be tested directly using an F-test or a Likelihood Ratio test. In this thesis, we consider a different approach to testing joint hypotheses: taking the p-values $(P_j)_{j\in\mathcal{J}}$ for the marginal hypotheses $(H_j)_{j\in\mathcal{J}}$, we aim to construct a function $f : (0,1)^{|\mathcal{J}|} \to (0,1)$ which combines these p-values, forming a test of $H_{\mathcal{J}}$. This approach is referred to as combination testing. Methods for combining p-values date back to Fisher, and since then, many combination tests have been proposed. These combination tests can be of great interest in closed testing, but the testing of a joint hypothesis can also be useful in and of itself. For example, there may be multiple, seemingly conflicting, studies about the effect of a drug. With combination testing, we can combine these studies and ask the question: 'was there any effect in at least one of the studies?'.

In Paper **A**, we propose a new family of combination tests based on the order statistics of independent p-values, and we argue that this family of combination tests has high power in many scenarios.

1.3 Causal discovery

Statistical learning deals with the problem of learning properties of distributions of random variables – say $(X_i, Y_i)_{i=1}^n$ – from observed outcomes of these variables, $(x_i, y_i)_{i=1}^n$. We may, for instance, be interested in predicting the value of Y from a new observation of X, i.e., the conditional mean of Y given X. This can be done by imposing additional assumptions on the joint distribution of (X_i, Y_i) – that is, by assuming a statistical model.[4] For instance, we may assume that each pair (X_i, Y_i) is sampled independently from one

[4]Formally, a statistical model for (X, Y) consists of a measurable space and a family of probability measures on this measurable space.

another, and that X_i and Y_i are linearly related with Gaussian errors,

$$Y_i = \beta X_i + N_i, \quad N_1, \ldots, N_n \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

In the above statistical model, the coefficient β can be consistently estimated (e.g., by Ordinary Least Squares estimation) and used to generate predictions $\mathbb{E}[Y \mid X = x] = \beta x$. The statistical model, however, cannot be used to draw inference about the causal relationship between X and Y. For instance, this model cannot be used to determine what happens to Y if X is intervened upon – that is, if we actively set the value of X, what happens to Y? To answer this question, we need a causal model.

Causal models are enhanced statistical models. Like a statistical model, a causal model specifies a family of possible joint distributions of a set of random variables. In addition, a causal model also specifies how the system acts under external manipulation. Thus, a causal model contains a statistical model.

In this thesis, we use the framework of Structural Causal Models[5] (SCMs) (Pearl, 2009; Peters et al., 2017). We formally define SCMs in Definition 1.10. Briefly, an SCM consists of a set of structural assignments and a noise distribution. Furthermore, an SCM has an accompanying graph, which we use to visualize the causal structure of a system. In this thesis, we are primarily concerned with the task of learning the graphs of these models from data. In Paper **B**, we consider the task of learning summary graphs (see Section 1.3.6) from time series data, and in Paper **C** we attempt to learn particular subgraphs of interest. To facilitate the discussion of these tasks, we briefly introduce graphs, SCMs, and key concepts in causality in the following sections.

1.3.1 Graphs

A directed graph $\mathcal{G} \coloneqq (V, E)$ consists of a set of vertices V and directed edges $E \subseteq \{(a \to b) \mid a, b \in V\}$. A walk ϵ is a sequence of edges. If ϵ has no repeated edges, we say that ϵ is a path. A cycle is a path that starts and ends in the same vertex. If all edges on a path point in the same direction, we call it a directed path. We say that $\mathcal{G}$ is a Directed Acyclic Graph (DAG) if $\mathcal{G}$ has no directed cycles.

The parents of a vertex Y, denoted PA_Y, is the set of vertices from which there exists an edge with endpoint in Y. The ancestors of Y, denoted AN_Y, are the vertices from which a directed path to Y exists. Similarly, the children CH_Y and descendants DE_Y of Y are the vertices to which a directed edge or path exists from Y, respectively. The non-descendants ND_Y of Y are the vertices in V that are not in DE_Y (not including Y itself).

An important tool in analyzing graphs – and thereby causal relationships in an SCM – is that of d-separation (Pearl, 2009).

[5] Also known as Structural Equation Models.

Definition 1.8 (Pearl, 2009)**.** *Let $\mathcal{G} = (V, E)$ be a DAG and let ϵ be a path in $\mathcal{G}$. Let $A, B, C \subseteq V$ be distinct subsets of vertices. The path ϵ is blocked by C if and only if at least one of the following two points holds:*

- *The path ϵ contains a directed path $a \to c \to b$ or fork $a \leftarrow c \to b$ such that $c \in C$.*
- *The path ϵ contains an inverted fork $a \to c \leftarrow b$ (with c being called a collider) such that $c \notin C$ and $\mathrm{DE}_c \cap C = \emptyset$.*

If all paths between A and B are blocked by C, we say that A and B are d-separated by C and write this as $A \perp_d B \mid C$.

1.3.2 Conditional independence and Markov properties

Graphs serve two overall purposes in this thesis: 1) through d-separation, they provide a tool for proving statements about conditional independence, and 2) they are useful for visualizing potential causal relationships in a system. For graphs to serve these two purposes, they need to act as graphical representations of the random variables of interest. That is, there needs to be a correspondence between d-separation in some graph $\mathcal{G} = (V, E)$ and conditional independence statements between a collection of random variables $(X_v)_{v \in V}$.

Definition 1.9 (see, e.g., Peters et al., 2017)**.** *Let $\mathcal{G} = (V, E)$ be a graph, let $(X_v)_{v \in V}$ be random variables and let $\mathbb{P}_X$ be the joint distribution of $(X_v)_{v \in V}$. We say that $\mathbb{P}_X$ obeys the Directed Global Markov (DGM) property relative to $\mathcal{G}$ if every d-separation in $\mathcal{G}$ implies a conditional independence in $(X_v)_{v \in V}$, i.e., if for all disjoint sets $A, B, C \subseteq V$ it holds that*

$$A \perp_d B \mid C \implies X_A \perp\!\!\!\perp X_B \mid X_C.$$

If the above implication is a bi-implication, we say that $\mathbb{P}_X$ is faithful to $\mathcal{G}$.

It is trivial to construct a graph $\mathcal{G}$ such that $\mathbb{P}_X$ obeys DGM relative to $\mathcal{G}$. For instance, any fully connected graph with vertex set V will satisfy this trivially, as there are no d-separations. Thus, for a graphical representation of $(X_v)_{v \in V}$ to be useful, it has to be sufficiently sparse.

1.3.3 Structural Causal Models

We introduce here the concept of Structural Causal Models (SCMs).

Definition 1.10 (see, e.g., Peters et al., 2017)**.** *A structural causal model $\mathcal{C} = (\mathcal{S}, \mathbb{Q})$ consists of a noise distribution $\mathbb{Q}$ on $\mathbb{R}^d$ with mutually independent marginals and a set $\mathcal{S}$ of d structural assignments,*

$$\forall v \in [d]: \quad X_v \coloneqq f_v(X_{\mathrm{PA}_v}, N_v), \tag{1.2}$$

where $(N_1, \dots, N_d) \sim \mathbb{Q}$.

The causal graph of $\mathcal{C}$ *is obtained by letting* $\mathcal{G} = (V, E)$ *with* $V = [d]$ *and*

$$E = \{(a \to b) \mid X_a \text{ enters the structural assignment of } X_b\}.$$

The causal order of $\mathcal{C}$*, if it exists, is a permutation* $\pi : [d] \to [d]$ *such that* $a \in \mathrm{PA}_b$ *implies* $\pi(a) < \pi(b)$*. If a causal ordering of* $\mathcal{C}$ *exists, we say that* $\mathcal{C}$ *is acyclic. If not, we say that* $\mathcal{C}$ *is cyclic.*

Whether a solution to an SCM $\mathcal{C}$ exists (i.e., whether there exists $X = (X_v)_{v \in V}$ with $N = (N_v)_{v \in V} \sim \mathbb{Q}$ such that $X \overset{a.s.}{=} (f_v(X_{\mathrm{PA}_v}, N_v))_{v \in V}$) depends on whether $\mathcal{C}$ is acyclic or not. If $\mathcal{C}$ is acyclic, a solution exists and $\mathcal{C}$ implies a unique joint distribution such that $(X_v)_{v \in V} \sim \mathbb{P}_X$ (Peters et al., 2017). If $\mathcal{C}$ is cyclic, there does not necessarily exist a solution (for more information on solutions to cyclic SCMs, see Bongers et al., 2021). In this thesis, we consider only acyclic SCMs.

An attractive property of SCMs is that their implied distributions obey a Markov property relative to their causal graphs.

Theorem 1.1 (Pearl, 2009, Theorem 1.4.1)**.** *Let* $\mathcal{C}$ *be an acyclic SCM. The implied distribution of* $\mathcal{C}$ *satisfies the Directed Local Markov (DLM) property relative to the causal graph of* $\mathcal{C}$*, i.e.,*

$$\forall v \in V : \quad X_v \perp\!\!\!\perp X_{\mathrm{ND}_v \setminus \mathrm{PA}_v} \mid X_{\mathrm{PA}_v}.$$

Combined with the following theorem, we see that if the implied distribution of an SCM is dominated by a product measure, then it obeys DGM relative to its causal graph.

Theorem 1.2 (Lauritzen, 1996, Theorem 3.27)**.** *Let* $\mathcal{G}$ *be a DAG and* $\mathbb{P}_X$ *be a distribution that has density with respect to a product measure* μ*. The following statements are equivalent.*

- $\mathbb{P}_X$ *factorizes with respect to* $\mathcal{G}$*.*
- $\mathbb{P}_X$ *obeys DLM relative to* $\mathcal{G}$*.*
- $\mathbb{P}_X$ *obeys DGM relative to* $\mathcal{G}$*.*

Throughout this thesis, we implicitly assume that the implied distribution of any considered SCM is dominated by a product measure (i.e., it has density). The factorization property in the above theorem is useful, as it decomposes the problem of making inference about the joint distribution $\mathbb{P}_X$ into a series of (potentially) less complicated sub-problems, namely making inference about the conditionals of the marginals. However, we shall not make explicit use of the factorization property in this thesis, and therefore we do not elaborate on it further.

1.3.4 Interventions and causal effects

Thus far, we have not gone beyond the properties of usual statistical models in our discussion of SCMs. The key property of causal models is that we can make statements about what happens when we intervene in the system.

Definition 1.11 (see, e.g., Peters et al., 2017)**.** *Let $\mathcal{G}$ be the graph of an SCM $\mathcal{C} = (\mathcal{S}, \mathbb{Q})$ with implied distribution $\mathbb{P}_X$. An intervention to $\mathcal{C}$ consists of replacing some (or all) of the structural equations $\mathcal{S}$ with new:*

$$\forall v \text{ that are intervened on}: \quad X_v \coloneqq \widetilde{f}_v(X_{\tilde{\mathrm{PA}}_v}, \tilde{N}_v).$$

We write the intervened upon model $\tilde{\mathcal{C}} \coloneqq (\tilde{\mathcal{S}}, \tilde{\mathbb{Q}})$. We call the implied distribution of $\tilde{\mathcal{C}}$ the interventional distribution and denote it by $\mathbb{P}_X^{\mathrm{do}(\tilde{S})}$. We denote the marginal distribution of an intervened upon variable X_v by $\mathbb{P}_{X_v}^{\mathrm{do}(X_v \coloneqq \tilde{f}_v(X_{\tilde{\mathrm{PA}}_v}, \tilde{N}_v))}$.

An atomic intervention consists of setting a variable X_v to a fixed value $c \in \mathbb{R}$. Thus, an atomic intervention consists of removing all edges going into v in the graph $\mathcal{G}$, but leaving all outgoing edges, and setting $X_v \coloneqq c$. Interventions do not need to be atomic. In fact, they do not even need to remove any incoming edges, but can simply consist of modifications to the causal function f_v. When we talk of interventions in Paper **C**, we do not make any assumptions about whether these are atomic or not.

Definition 1.12 (Peters et al., 2017)**.** *Let $\mathcal{G}$ be the graph of an SCM $\mathcal{C} = (\mathcal{S}, \mathbb{Q})$ with implied distribution $\mathbb{P}_X$. Fix two distinct labels a, b in the graph $\mathcal{G}$. We say that there is a total causal effect from a to b (or that a causes b) if and only if*

$$\exists c \in \mathbb{R}: \quad \mathbb{P}_{X_b}^{\mathrm{do}(X_a \coloneqq c)} \neq \mathbb{P}_{X_b}.$$

That is, there is a total causal effect from a to b if there exists an atomic intervention on X_a that renders the interventional distribution of X_b different from its observational distribution.

It is important to note that the existence of a directed path from a to b does not guarantee that a causes b. However, it is necessary that a directed path from a to b exists, for there to be a total causal effect from a to b (Peters et al., 2017). Thus, when we discuss the learning of graphs in Chapters 3 and 4, we cannot use the learned graphs to say anything about the causal relations in the investigated systems without further assumptions. However, we can use them to say something about the potential causal effects.

1.3.5 Invariance and Invariant Causal Prediction

Recently, Peters et al. (2016) introduced Invariant Causal Prediction (ICP), which exploits the invariance of causal models with respect to changing environments in order to learn the parents of a response variable. This paper is

a cornerstone of Paper **C**, and we therefore briefly introduce the concepts of invariance and ICP here.

Suppose that we observe a response variable Y and d predictor variables $X = (X_1, \ldots, X_d)$, and that we observe these variables across different environments E. This can, for example, be data collected from an observational setting and an interventional setting. The key observation of ICP is that any causal model will be invariant to changes in the environment. That is, a model $X_S = (X_s)_{s \in S}$ is invariant if it renders the response independent of the environment in the conditional distribution of X_S, i.e., $Y \perp\!\!\!\perp E \mid X_S$. Since any causal model must be invariant, ICP tests all possible subsets of predictors for invariance and outputs the intersection of all sets where the hypothesis of invariance was not rejected. Peters et al. (2016) show that the oracle output of ICP contains only parents of the response and that it can be learned from data with high probability.

The notion of invariance has (at least) a two-fold importance. First, learning the parents of a response variable gives us valuable insight into the causal structure of an observed system, and we can use this insight to identify possible intervention targets. Second, models that are invariant to changes in the environment produce invariant predictions. That is, the prediction of Y, based on a new set of observed predictor variables, will have the same error distribution no matter the environment it is predicted in.

Still, ICP can fail in both of the above regards: if multiple, disjoint models are invariant, ICP fails to output anything. In addition, even when ICP outputs a non-empty set, there is no guarantee that this set is invariant. These two observations were the motivation for Paper **C**, where we use the concept of *minimal invariance* (see Paper **C**, Definition 3.1), to modify the ICP procedure, such that the output is (under mild assumptions) always non-empty and invariant. This improvement does come at a cost: the statistical guarantees of ICP hold for any finite sample, while the corresponding guarantees we make in Paper **C** are asymptotic. In addition, ICP outputs only parents of the response, while the method we propose in Paper **C** outputs ancestors of the response. In Section 4.2, we sketch a method to post hoc filter the parents from the non-parental ancestors output by the proposed method.

1.3.6 Summary graphs of infinite-dimensional causal graphs

In our above description of SCMs and causal notions, we have implicitly assumed that all systems are time-independent. When causal systems are time-dependent, there are several different approaches to causal learning. Peters et al. (2022) propose an extension of SCMs – called Causal Kinetic Models – in which the structural assignments of an SCM are replaced with a set of stochastic differential equations. Schweder (1970) proposes the concept of Conditional Local Independence (CLI), which formalizes the notion of whether the past of one continuous-time stochastic process influences another. Building on CLI,

Didelez (2007, 2008, 2015) has proposed the use of local independence graphs, and several methods for learning these have been proposed (see, e.g., Mogensen et al., 2018; Mogensen, 2020; Mogensen and Hansen, 2020).

In Paper **B**, we take a different and more simplistic approach to learning the causal relations of a time-dependent system. Here, we consider discrete-time stochastic processes that are time-homogeneous. These processes have infinite-dimensional causal graphs and we attempt to learn finite-dimensional summary graphs of these.

Formally, We consider a discrete-time stochastic process $(X(t))_{t\in\mathbb{N}_0}$ where each $X(t)$ is a d-valued random variable on a probability space $(\mathcal{X}^d, \mathcal{F}, \mathcal{P})$. Furthermore, we let $(N(t))_{t\in\mathbb{N}_0}$ denote an infinite sequence of mutually independent noise innovations. We assume a time-homogeneous causal structure of this process, i.e., that every $X(t)$ is structurally generated by its past values and that there are no instantaneous effects. In other words, we assume the existence of a causal function F such that $X(t)$ is structurally assigned as

$$X(t) \coloneqq F(X(t-l), \dots, X(t-1)) + N(t),$$

where l is the (unknown) number of lags in the time series. The causal graph $\mathcal{G}^\infty$ of this system is infinitely large. For such infinite-dimensional systems, the causal graph is difficult, or impossible even, to comprehend. Thus, we define instead the *summary graph* of a discrete-time stochastic process as a directed graph $\mathcal{G} = (V, E)$ with d nodes $V = [d]$ and edges

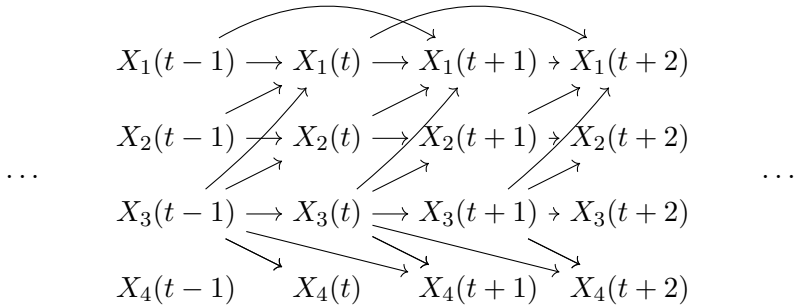
$$E = \{(i \to j) \mid \exists t_1 < t_2 \text{ s.t. } X_i(t_1) \text{ is in the structural assignment of } X_j(t_2)\}.$$

That is, the summary graph contains only d nodes and has an edge from X_i to X_j if and only if the past of X_i enters the structural assignment of X_j at some point in time. Thus, the summary graph encodes the Granger-causes (Granger, 1969) of each marginal process $X_i(t)$. Note that the summary graph does not constitute a graphical model in the classical sense, because it cannot be used to read off conditional independencies using Markov properties. See Figure 1.1a for an example of a causal graph of a time series and Figure 1.1b for its corresponding summary graph.

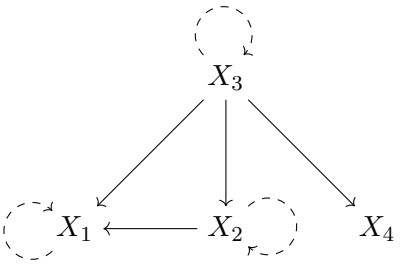
In Paper **B**, we take the following approach to learning summary graphs. For each marginal process $X_j(t)$, we consider the expected absolute value of the partial derivative of its causal function F_j (assuming that it is differentiable) with respect to another marginal process X_i evaluated at time t':

$$\theta_{ij}^t = \mathbb{E}|\partial_i F_j(X(t'))|.$$

The intuition is, that this parameter captures the expected effect of $X_i(t')$ on $X_j(t)$. If $X_j(t)$ does not functionally depend on $X_i(t')$, the partial derivative is zero and thus $\theta_{ij}^t = 0$. On the other hand, if there is a functional dependence between the two variables, the partial derivative will be non-zero somewhere



(a) The graph of an infinite discrete-time stochastic process.



(b) The corresponding summary graph. A self-cycle represents a lagged time effect of a node onto itself.

Figure 1.1: An example of an infinite-dimensional graph and its corresponding summary graph.

and $\theta_{ij}^t > 0$. In Paper **B**, we approximate θ_{ij}^t by employing subsampling and linear regression methods to detect non-constant regions of F_j. However, we do not assume linearity of F_j itself. The intuition behind this approach is, that even if F is non-linear, it may appear locally linear in some regions, which can be detected by linear regression methods. In addition, we discuss a simulation artifact in Additive Noise Models that can render the causal ordering identified by marginal variances. Lastly, we discuss how the presence of this artifact affects structure learning algorithms.

Chapter 2

Multiple testing

In this chapter, we present the paper 'Too Many, Too Improbable: testing joint hypotheses and closed testing shortcuts'. This paper has been submitted to the Journal of Statistical Planning and Inference and is still under review at the time of writing. In Paper **A**, we propose a new family of combination tests and show that these are contained in a larger family that allow for shortcuts in CTPs. In Section 2.2 we elaborate on the intuition and motivation for the proposed test statistics, and in Section 2.4 we give examples of how one shows that a test procedure satisfies the shortcuts of Paper **A**. In Section 2.5, we argue that p-values for the proposed combination tests can be easily approximated, yielding considerable computational speedups. In Sections 2.7.1 and 2.7.2, we apply our methods to real data, and finally, in Section 2.8 we describe the `R` package `TMTI` (Mogensen, 2021), which was created to accompany Paper **A**.

2.1 Paper A

Phillip B. Mogensen and Bo Markussen. 'Too Many, Too Improbable: testing joint hypotheses and closed testing shortcuts' arXiv preprint arXiv:2108.04731 (2021).

Too Many, Too Improbable: testing joint hypotheses and closed testing shortcuts

Phillip B. Mogensen and Bo Markussen[1]

[1]Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, Copenhagen, Denmark

Abstract

Hypothesis testing is a key part of empirical science and multiple testing as well as the combination of evidence from several tests are continued areas of research. In this article we consider the problem of combining the results of multiple hypothesis tests to i) test global hypotheses and ii) make marginal inference while controlling the k-FWER. We propose a new family of combination tests for joint hypotheses, called the 'Too Many, Too Improbable' (TMTI) statistics, which we show through simulation to have higher power than other combination tests against many alternatives. Furthermore, we prove that a large family of combination tests – which includes the one we propose but also other combination tests – admits a quadratic shortcut when used in a Closed Testing Procedure, which controls the FWER strongly. We develop an algorithm that is linear in the number of hypotheses for obtaining confidence sets for the number of false hypotheses among a collection of hypotheses and an algorithm that is cubic in the number of hypotheses for controlling the k-FWER for any k greater than one.

1 Introduction

The problem of combining p-values from tests of a family of hypotheses $\{H_i\}_{i\in\mathcal{I}}$ indexed by a finite set $\mathcal{I}$ has long been a field of study and remains an active area of research today. In 1925, Fisher proposed a method of combining independent p-values by observing that minus two times the sum of $(\log p_i)_{i\in\mathcal{I}}$ follows a $\chi^2_{2|\mathcal{I}|}$-distribution (Fisher, 1992). Fisher's combination test is asymptotically Bahadur-optimal among the class of all combination tests (Littell and Folks, 1973). Still, the Fisher Combination Test can potentially be outperformed by other combination tests for any given finite sample. In 1973, Brown devised an extension of the Fisher Combination Test or when the underlying test statistics are jointly Gaussian with a known covariance matrix and the hypotheses are one-tailed (Brown, 1975). Kost and McDermott (2002) further relaxed the assumptions on the dependence structure by deriving an approximation of the distribution of the Fisher Combination Test when the underlying tests statistics are jointly T-distributed with a common denominator. In recent years the Cauchy Combination Test (Liu and Xie, 2020) and the Harmonic Mean p-value (Wilson, 2019) have been proposed and Vovk and Wang (2020) derive a large family of combination tests by using the Kolmogorov generalized f-mean.

These combination-based methods for testing the global null hypothesis $H_0 \coloneqq \bigcap_{i\in\mathcal{I}} H_i$ follow the overall recipe of finding a mapping $f : [0,1]^{|\mathcal{I}|} \to [0,1]$ of $\boldsymbol{p} \coloneqq (p_i)_{i\in\mathcal{I}}$ such that $f(\boldsymbol{p})$ is again a p-value under H_0. That is, such that for any choice of $\alpha \in [0,1]$ it holds that $\mathbb{P}(f(\boldsymbol{p}) \leq \alpha) \leq \alpha$ when H_0 is true.

One particular way of obtaining this property is to choose *any* function, say f_1, that maps the hypercube $[0,1]^{|\mathcal{I}|}$ to any subset of the real line and then subsequently transform the resulting random variable by its cumulative distribution function (CDF), say f_2. The composite mapping $f_2 \circ f_1$ is then a valid combination test. The Fisher Combination Test is an example of this; first, we map $\boldsymbol{p}$ onto the positive real line by the mapping $\boldsymbol{p} \mapsto -2\sum_{i\in\mathcal{I}} \log p_i$, which is then transformed back onto the unit interval using the CDF of a $\chi^2_{2|\mathcal{I}|}$-distribution. Another simple way of constructing valid combination tests is to use the minimal p-value from any procedure that controls the family-wise error rate (FWER). For example, we may use the minimal p-value of the Bonferroni corrected p-values, corresponding to the mapping $\boldsymbol{p} \mapsto \min\left(|\mathcal{I}| \cdot \min(\boldsymbol{p}), 1\right)$.

In this paper, we introduce a family of combination tests – the 'Too Many, Too Improbable' (TMTI) tests – that strongly controls the Type I error at level α, for any choice of $\alpha \in (0,1)$. In brief, these statistics are obtained by ordering the observed p-values, transforming them by the CDFs of beta distributions and returning a local minimum. The p-value is then the local minimum transformed by its CDF. We derive analytical expressions for the null CDFs of the TMTI test statistics under an assumption of independence and show through simulation that the TMTI tests can have higher power than other common combination tests under many alternatives. Additionally, we give an $\mathcal{O}(m^2)$ shortcut for carrying out a full Closed Testing Procedure for all elementary hypotheses for a large family of test statistics, obtained by considering test statistics of the form $Z = h(F_{(1)}(p_{(1)}), \ldots, F_{(m)}(p_{(m)}))$ under mild assumptions on the functions $F_{(1)}, \ldots, F_{(m)}$ and h. Using prior work by Goeman and Solari (2011), we show how these shortcuts can be used to obtain k-FWER control for elementary hypotheses as well as construct confidence sets for the number of false hypotheses in a rejection set. Finally, we discuss how mixing different local tests across a Closed Testing Procedure can be used to increase power.

2 The 'Too Many, Too Improbable' family of test statistics

2.1 Notation and setup

Let $\mathcal{I} \coloneqq \{1, \ldots, m\}$ be an index set with cardinality m and let $\{H_i\}_{i\in\mathcal{I}}$ be hypotheses. Let $(P_i)_{i\in\mathcal{I}}$ be random variables on probability spaces $(\Omega_i, \mathbb{B}_i, \mathbb{P}_i)_{i\in\mathcal{I}}$ with $\Omega_i \subseteq [0,1]$. In most situations we will have $\Omega_i = [0,1]$ and have $\mathbb{B}_i$ be the Borel sigma-algebra, although this need not be the case. We denote by p_i an outcome of P_i and call p_i the p-value for the test of H_i. For a given subset of indices, $\mathcal{J} \subseteq \mathcal{I}$, we consider the task of testing the joint hypothesis $H_\mathcal{J} \coloneqq \bigcap_{j\in\mathcal{J}} H_j$ by using the marginal p-values, $\boldsymbol{P}^\mathcal{J} \coloneqq (P_j)_{j\in\mathcal{J}}$. The set $\mathcal{J}$ can be chosen freely according to what kind of hypothesis one wishes to test. If we choose $\mathcal{J}$ with $|\mathcal{J}| = 1$, no adjustment needs to be made, as we are simply testing a marginal hypothesis, for which we already have a p-value. If we choose $\mathcal{J} = \mathcal{I}$, we are considering the global null hypothesis of $\mathcal{I}$. Anything in between those two extremes corresponds to testing a particular joint hypothesis. E.g., if $(p_i)_{i\in\mathcal{I}}$ are the p-values output from a genome-wide association study, then $\mathcal{J}$ could correspond to a particular region, which is of special interest, e.g., a gene or chromosome.

In order to test $H_\mathcal{J}$, we construct a test statistic, denoted by Z, with corresponding p-value P that satisfies

$$H_\mathcal{J} \text{ true } \implies \forall \alpha \in [0,1] : \mathbb{P}(P \leq \alpha) \leq \alpha. \tag{1}$$

The above statement is called Type I error control and means that whenever the joint hypothesis $H_\mathcal{J}$ is true, the probability that we reject $H_\mathcal{J}$ at level α is at most α.

2.2 Definition of the TMTI statistics

Let $P^{\mathcal{J}}_{(1)} \leq \ldots, P^{\mathcal{J}}_{(|\mathcal{J}|)}$ denote an ordering of $\boldsymbol{P}^{\mathcal{J}}$. This ordering is possibly not unique. Let $\beta(a,b)(x)$ denote the CDF of the $\beta(a,b)$-distribution.When the shape and scale parameters are integers, we can write

$$\forall i, m \in \mathbb{N}: \qquad \beta(i, m+1-i)(x) = \sum_{k=i}^{m} \binom{m}{k} x^k (1-x)^{m-k}. \tag{2}$$

We construct the collection $\boldsymbol{Y}^{\mathcal{J}} \coloneqq (Y^{\mathcal{J}}_k)_{k=0}^{|\mathcal{J}|+1}$ of random variables by $Y^{\mathcal{J}}_0 \coloneqq 2$, $Y^{\mathcal{J}}_{|\mathcal{J}|+1} \coloneqq 2$, and

$$\forall k \in \{1, \ldots, |\mathcal{J}|\}: \qquad Y^{\mathcal{J}}_k \coloneqq \beta(k, |\mathcal{J}|+1-k)(P^{\mathcal{J}}_{(k)}).$$

If all variables in $\boldsymbol{P}^{\mathcal{J}}$ are independent and exactly uniform, then each $Y^{\mathcal{J}}_k$ is uniformly distributed on $[0,1]$ for $k = 1, \ldots, |\mathcal{J}|$, as it is well known that the order statistics of i.i.d. $U(0,1)$ variables are β-distributed.

Let $c \leq |\mathcal{J}|$ be an integer. We then consider the first $Y^{\mathcal{J}}_k$ among the first c variables that is strictly smaller than the following $n \in \mathbb{N} \cup \{\infty\}$, i.e.,

$$L_{n,c} \coloneqq \min\{l \in \{1, \ldots, c\} : Y^{\mathcal{J}}_l < Y^{\mathcal{J}}_k \text{ for all } k = l+1, \ldots, \min(l+n, |\mathcal{J}|+1)\}.$$

If $Y^{\mathcal{J}}_1 \geq \cdots \geq Y^{\mathcal{J}}_c$ we set $L_{n,c} = c$. We think of $L_{n,c}$ as the index of a kind of local minimum of $Y^{\mathcal{J}}_1, \ldots, Y^{\mathcal{J}}_c$, in the sense that $Y^{\mathcal{J}}_{L_{n,c}}$ is always a local minimum, but it further needs to satisfy that it is smaller than the following n terms. In particular, $Y^{\mathcal{J}}_{L_{1,c}}$ is the first local minimum of $Y^{\mathcal{J}}_1, \ldots, Y^{\mathcal{J}}_c$ and $Y^{\mathcal{J}}_{L_{\infty,c}}$ is the global minimum of $Y^{\mathcal{J}}_1, \ldots, Y^{\mathcal{J}}_c$. The construction of $Y^{\mathcal{J}}_0$ and $Y^{\mathcal{J}}_{|\mathcal{J}|+1}$ is a technical one, meant only to ensure the existence of $L^{\mathcal{J}}_{n,c}$. To ease the notational burden, we omit the subscripted n and c and the superscripted $\mathcal{J}$ when the particular choices of n, c and $\mathcal{J}$ are not of importance or unambiguous from context.

Definition 1. *Let $n \in \mathbb{N} \cup \{\infty\}$ and let $c \leq |\mathcal{J}|$ be an integer. The 'Too Many, Too Improbable' test statistic is then defined as*

$$Z^{\mathcal{J}}_{n,c} \coloneqq Y^{\mathcal{J}}_{L_{n,c}}.$$

Small values of $Z^{\mathcal{J}}_{n,c}$ are critical and the p-value for the test of $H_{\mathcal{J}}$ is obtained by evaluating the test statistic in its CDF under $H_{\mathcal{J}}$. We denote by $\gamma^{\mathcal{J}}_{n,c}(x)$ the CDF of $Z^{\mathcal{J}}_{n,c}$ under $H_{\mathcal{J}}$.

Generally, we will only consider cases in which $n = 1$ or $n = \infty$, as these are the most natural choices of n. However, the setup allows for other choices of n. Choosing $1 < n < \infty$ can potentially increase the power of the procedure in cases where signals are fairly sparse, but sufficiently weak that the first local minimum falls 'too early' by chance. However, we do not investigate this further, but simply remark that it is possible to choose n different from what we consider in the remainder of this paper.

Remark 1. *Testing the joint hypothesis $H_{\mathcal{J}}$ using any TMTI test satisfies the statement in* (1) *by the probability integral transform. That is, the TMTI test controls the Type I error.*

Remark 2. *Whenever $|\mathcal{J}| = 1$, say $\mathcal{J} = \{j\}$, the TMTI transform is simply the identity transform, i.e., $\gamma(Z) = P_j$.*

Remark 3. *If the variables in $\boldsymbol{P}^{\mathcal{I}}$ are exchangeable, i.e., if any two subsets of equal size have the same joint distribution, it follows, that for any two sets $\mathcal{J}_1, \mathcal{J}_2 \subset \mathcal{I}$ with $|\mathcal{J}_1| = |\mathcal{J}_2|$, we have $\gamma^{\mathcal{J}_1} = \gamma^{\mathcal{J}_2}$. Thus, for exchangeable p-values, the CDF of the TMTI statistic depends only on the choice of $\mathcal{J}$ through its cardinality.*

2.3 Truncation procedures and the TMTI

The Truncated Product Method of Zaykin et al. (2002) and the Rank Truncated Product Method of Dudbridge and Koeleman (2003) are two notable variants of the Fisher Combination Test, that also test the global null hypothesis but against different alternatives.

The Truncated Product Method is a combination test that uses only the p-values that are smaller than some predefined threshold $\tau \in (0, 1)$. The alternative hypothesis is therefore, that there is at least one false hypothesis among those hypotheses that gave rise to p-values below τ. The Rank Truncated Product Method is also a combination test, but this uses only the smallest K p-values, for some predefined $K < |\mathcal{J}|$. Thus, the alternative hypothesis is, that there is at least one false hypothesis among those, that gave rise to the K smallest p-values.

The TMTI family of test statistics includes similar procedures. For any $c < |\mathcal{J}|$, the alternative hypothesis is that there is at least one false hypothesis among those that gave rise to the c smallest p-values. Thus, setting $c = K$ for some integer $K < |\mathcal{J}|$, the TMTI procedure uses only the first K p-values in the construction of the test statistic and therefore tests the joint hypothesis $H_{\mathcal{J}}$ against the same alternative as the Rank Truncated Product Method. We call this procedure the rank truncated TMTI.

By setting $c = \max(\{j \in \{1, \ldots, |\mathcal{J}|\} : p_{(j)} \leq \tau\} \cup \{1\}) =: \bar{\tau}$, for some value $\tau \in (0, 1)$, the TMTI procedure uses only the p-values that are marginally significant at level τ, and thus tests against the same alternative as the Truncated Product Method. We call this procedure the truncated TMTI. In the event that no p-values are smaller than τ, c becomes 1 and uses instead the smallest of the available p-values.

We write TMTI_n to denote the TMTI statistic $Z^{\mathcal{J}}_{n,c}$ with $c = |\mathcal{J}|$, tTMTI_n to denote the truncated TMTI statistic and rtTMTI_n to denote the rank truncated TMTI statistic.

There are two potential advantages to using a truncated procedure (i.e., $c < |\mathcal{J}|$) over a non-truncated procedure. First, for large m (say, $m \geq 10^6$), it is non-trivial to compute the TMTI_∞-statistic, because its computation involves sorting m different p-values and computing m different β-transformations. Using a truncation procedure instead reduces the computational cost, because fewer p-values need to be considered. Thus, only a partial sorting is required and fewer β-transformations need to be computed. Second, as we outline below, using a truncation procedure can potentially have higher power than its non-truncated version.

Lemma 1. *Let $n \in \mathbb{N} \cup \{\infty\}$ and $x \in (0, 1)$. Let $\mathcal{I}$ be an index set with cardinality m. It follows that*

$$\forall c < m : \qquad \gamma_{n,c}(x) < \gamma_{n,m}(x)$$

When using $n = 1$, i.e., considering the first local minimum, and when using moderate values of τ and K, it is likely that $\gamma_{1,\tau}$, $\gamma_{1,K}$ and γ_1 are going to be nearly identical, as the first local minimum is likely to lie early in the sequence $\boldsymbol{Y}$. This implies that the p-values of the tTMTI_1, the rtTMTI_1 and TMTI_1 tests are nearly identical. Thus, the TMTI_1 by itself can be thought of as a truncation method. However, if using the global minimum, we expect that there can be a large difference between the methods, as the global minimum is likely to lie further along the sequence $\boldsymbol{Y}$. Thus, applying either the tTMTI_∞ with a low τ or rtTMTI_∞ with a low K is going to be roughly equivalent to applying the TMTI_1. These properties are demonstrated in Figure 1 for the case of independent and exactly uniform p-values.

From Figure 1 we conclude, that if the global null hypothesis is indeed false, and if the global minimum of the sequence $\boldsymbol{Y}$ happens to fall within the first K or $\bar{\tau}$ indices of $\boldsymbol{Y}$, there is a potential for a large power gain by applying either the tTMTI_∞ or rtTMTI_∞ over the standard TMTI_∞. This is because the procedures will all be considering the same test statistic, Z, but they will evaluate it under different γ functions, thereby yielding different p-values. By Lemma 1, the γ functions from the tTMTI_∞ and rtTMTI_∞ procedures are dominated by the γ function from the TMTI_∞, implying that p-value resulting from applying the truncation procedures will be lower than those of the standard procedure, giving rise to higher power.

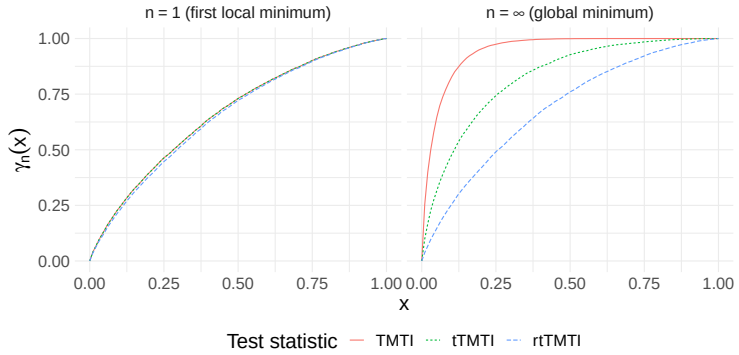


Figure 1: Comparison of $\gamma_n^{\mathcal{I}}$, $\gamma_{n,\tau}^{\mathcal{I}}$ and $\gamma_{n,K}^{\mathcal{I}}$ for $m = 10^5$, $\tau = 0.01$ and $K = 5$ in the case of independent and exactly uniform p-values. The solid red lines are TMTI, the dotted green and blue lines are tTMTI and rtTMTI, respectively.

3 Computation of γ

3.1 An analytical expression of the CDF of TMTI$_\infty$ in the i.i.d. case

In the case where the p-values are independent under the null hypothesis, we can derive an analytical expression of $\gamma_{\infty,c}$.

Theorem 1. *Let $P_1, \dots, P_m$ be i.i.d. uniformly distributed on $[0,1]$. For every $i \in \{1, \dots, m\}$, let x_i be the x-quantile of the $\beta(i, m+1-i)$ distribution and define the polynomial*

$$Q_i(x; \boldsymbol{a}) \coloneqq \sum_{j=1}^{i} \frac{a_j}{(i+1-j)!} x^{i+1-j}.$$

Define $\bar{Q}_1 \coloneqq x_1$ and define recursively

$$\forall i \in \{2, \dots, m\}: \qquad \bar{Q}_{i,c} \coloneqq Q_i(x_{\min(i,c)}; (1, -\bar{Q}_{1,c}, \dots, -\bar{Q}_{i-1,c})).$$

If $c \leq m$ is a fixed integer, then

$$\gamma_{\infty,c}(x) = x_c^m + \sum_{i=1}^{m-1} \frac{m!}{(m-i)!} \bar{Q}_{i,c}(1 - x_c^{m-i}). \tag{3}$$

Furthermore, let $\tau \in (0,1)$ and define $\tilde{Q}_1 \coloneqq \tau$ and

$$\forall i \in \{2, \dots, m\}: \qquad \tilde{Q}_i \coloneqq Q_i(\tau; (1, -\bar{Q}_{1,m}, \dots, \bar{Q}_{i-1,m})).$$

If c is a random variable given by $c = \max(\{i \in \{1, \dots, m\} : P_{(i)} < \tau\} \cup \{1\})$, then

$$\begin{aligned} \gamma_{\infty,c}(x) = {} & (1-\tau)^m \frac{x - \beta(1,m)(\tau)}{1 - \beta(1,m)(\tau)} I(x_1 > \tau) \\ & + \sum_{i=1}^{m} \left[\binom{m}{i} \tau^i (1-\tau)^{m-i} \left\{ 1 - \frac{i!}{\tau^i} \left(\tilde{Q}_i - \bar{Q}_{i,m} \right) I(x_i \leq \tau) \right\} \right] \end{aligned} \tag{4}$$

The above can be readily implemented by recursively computing the $\bar{Q}_{i,c}$ and $\tilde{Q}_i$ terms, e.g.,in a for-loop.

In the special case of $c = 1$, we have $\gamma_{n,1}(x) = x$ by construction, regardless of m and n. In this setting, the TMTI procedure is then a minimum-p method. This has the advantage, that the procedure can be applied directly in high-dimensional settings, if the assumption of independence holds. In particular, it is easy to show, that when $c = 1$ the critical value of the TMTI test is $1 - (1 - \alpha)^{1/m}$, and it is thus equivalent to using the Šidák correction (Šidák, 1967) for testing the global null hypothesis.

3.2 A bootstrap scheme for the CDF of TMTI$_n$ in the i.i.d. case

Although it is easy to implement Equations (3) and (4), numerical difficulties may arise when m is large, say $m > 100$, due to the presence of the factorials $1!, \ldots, m!$ in the computations. Essentially, the $\bar{Q}_{i,c}$ terms are all very small, because they include the reciprocals of factorials, but they are scaled up by another factorial. Although this is well-defined, numerical instabilities will occur in implementations in standard double-precision arithmetic. For larger m, one can perform the calculations in arbitrary precision, but the added computational cost of doing so can be enormous. Instead, a simple bootstrapping scheme can be employed by; i) drawing m values independently from a $U(0, 1)$ distribution; ii) transforming the values from step i as described in Equation 2.2 and saving the desired TMTI statistic as Z_b, where b indexes the iteration; iii) repeating steps i and ii sufficiently many times, say B, and; iv) using $\hat{\gamma}(x) \coloneqq \frac{1}{B}\sum_{b=1}^{B} I(Z_b \leq x)$ as an approximation of γ. This bootstrap scheme can be applied regardless of the choice of n and c.

3.3 The non-independent case

The derivation of the CDF, γ, in Section 3.1 relied on the assumption that all p-values are independent. Chen et al. (2020) argue that combining methods that are Valid under Arbitrary Dependence (VAD) structures have lower power than combining methods that are Valid under Independent (VI) p-values, if the underlying, p-values are in fact independent. However, if the underlying p-values are not independent, VI methods may fail to hold level, whereas their VAD counterparts will hold level for any dependence structure. Thus, the choice of combining method should depend on the scientific question of interest. For instance, in genome-wide association studies (GWAS), it is unreasonable to assume independence, as base-pairs are likely to be locally dependent (Dudbridge and Koeleman, 2003).[1]

In some cases, however, it is possible to apply the methods described in Sections 3.1 and 3.2, even if the p-values are not independent. For instance, if the underlying tests, Z, are jointly Gaussian with a known covariance matrix, Σ, these can be decorrelated by performing an eigendecomposition, $\Sigma = Q\Lambda Q^T$, and then constructing $\tilde{Z} \coloneqq (Q\Lambda^{-1/2}Q^T)^T Z$. Then, the components of $\tilde{Z}$ are jointly independent (see, e.g., Kessy et al. (2018)) and the methods described in Sections 3.1 and 3.2 can be directly applied to the p-values obtained from the test statistics $\tilde{Z}$. This remains true for any rotation, $\bar{Z} \coloneqq R(Q\Lambda^{-1/2}Q^T)^T Z$, where R is an orthogonal matrix.

If the p-values are not independent, and if the decorrelation procedure described above is not appropriate, one can still try to apply the TMTI directly. However, level of the test (i.e., Equation (1)) is no longer guaranteed, and thus there is a chance that the Type I error is increased. How much the Type I error increases depends entirely on the dependence structure of the p-values. In B.1, we investigate the level of the TMTI tests under three different types of dependencies: autoregressive p-values (i.e., $\mathrm{cor}(P_i, P_j) = \rho^{|i-j|}$), equicorrelated p-values (i.e., $\mathrm{cor}(P_i, P_j) = \rho$, for all i, j), and block-diagonally correlated p-values (i.e., $\mathrm{cor}(\boldsymbol{P})$ has a block-diagonal structure, where all off-diagonal entries are ρ if in the same block and 0 else).

[1] It is often possible to filter the p-values in a manner such that the remaining p-values are likely to be independent, e.g., using a distance-based filtering.

We note, however, that rtTMTI$_n$ seems to either have the correct level or be conservative, no matter the dependence structure. Overall, the TMTI tests hold level only under weak autoregressive and block-diagonal dependency structures, and fails to hold level for stronger dependency structures and equicorrelated p-values (see B.1 and Figures 5 and 6 for a full account of the results). Thus, the TMTI tests can potentially still be applied in settings with weak dependence, but it is not appropriate in settings with strong dependencies.

Finally, one can apply any VI combination method under arbitrary dependence, if one is able to sample from the joint distribution of $\boldsymbol{P}$ under the global null. This can, for instance, be done if one assumes an underlying parametric model or by employing a resampling bootstrap procedure. In B.2, we give an example of how this can be done in a case where the marginal hypotheses of interest are T-tests for the parameters in a linear model being zero.

4 Power of the TMTI – a simulation study

In this section, we show by means of simulation, that many of the TMTI tests have high power against a wide range of alternative hypotheses. In particular, we find that the TMTI$_\infty$ and tTMTI$_\infty$ tests have high power both in cases where signals are sparse but strong and in cases where signals are dense but weak.

We consider $m = 10^5$ independent tests, of which $N_{\text{false}} \in \{10^0, \cdots, 10^4\}$ are false. In order to investigate situations in which the p-values from true hypotheses are conservative, we generate these as $p_{\text{true}} := U^\delta$, where $U \sim U(0,1)$ and $\delta \in [0,1]$. When $\delta = 1$, this corresponds to the true p-values being exactly uniform, and when $\delta \in (0,1)$, it corresponds to the true p-values being strictly conservative. The degree of conservatism increases as δ decreases and the extreme case in which $\delta = 0$ corresponds to the degenerate case where all true p-values are equal to one, meaning that no hypothesis can ever be rejected, no matter the significance level. Situations in which the p-values are strictly conservative occur in many places. For instance, in a GWAS with dichotomous traits, the p-values will be conservative (Wu et al., 2011). Conservative p-values also occur in Invariant Causal Prediction, where a p-value for invariance is obtained as the minimum of Bonferroni-corrected p-values from multiple environments (Peters et al., 2016). We generate p-values for the false hypotheses by independently sampling Z-scores, $Z_1, \ldots, Z_{n_{\text{false}}} \overset{i.i.d.}{\sim} N(\mu_{\text{false}}, 1)$, for different values of μ_{false} and then letting $p_{i,\text{false}} := 2 \times (1 - \Phi(|Z_i|))$, where Φ is the CDF of a $N(0,1)$-distribution. The values of μ_{false} are chosen equidistantly between the two values, which satisfy that a Bonferroni test has either 5% or 99% power to reject the global null hypothesis in a setting with no conservatism.

For comparison, we include the Fisher Combination Test (which is known to lose power in the presence of conservative p-values (Zaykin et al., 2002)) and its truncated versions, and the Cauchy Combination Test and Harmonic Mean p-value (which are known to have high power in settings with sparse, strong signals (Liu and Xie, 2020; Wilson, 2019)). In all settings, we employ a significance level of $\alpha = 0.05$, and for the truncation procedures, we use $\tau = 0.05$ and $K = 10$. We include TMTI$_\infty$ and both truncation variants, as well as TMTI$_1$. For TMTI$_1$, we do not include any truncation variants, as we expect these to be roughly equal to the non-truncated version (per Figure 1). The results of the simulations are displayed in Figure 2. Overall, there are three things to notice.

First, the TMTI$_\infty$ and tTMTI$_\infty$ generally work well no matter how many false hypotheses there are. When there is only a single false hypothesis, these methods have less power than, e.g., a Bonferroni correction, which has the highest power in this scenario, but both methods have considerably higher power than both the Fisher Combination Test and Truncated Product Method. When there are more false hypotheses, TMTI$_\infty$ and tTMTI$_\infty$ perform on par with the Fisher Combination Test and Truncated Product Method, having considerably higher power than the remaining methods. No other methods exhibit this property; the

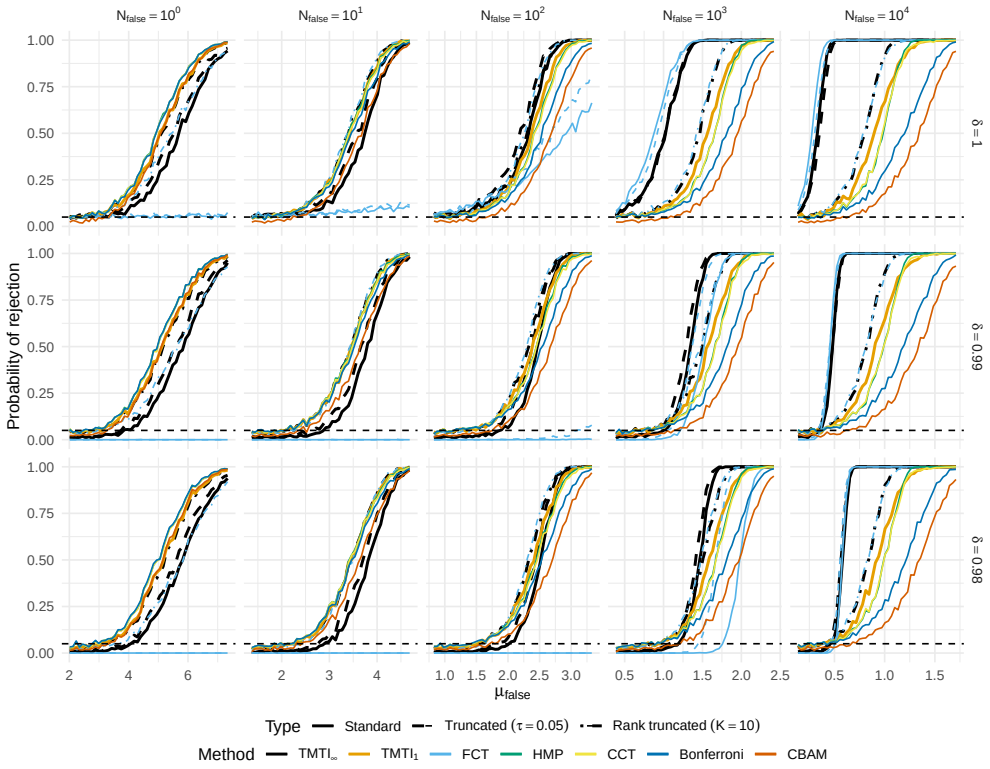


Figure 2: Power curves for different TMTI tests and competing methods. Generally, the $TMTI_\infty$ and $tTMTI_\infty$ work well in all settings. The values of μ_{false} are chosen equidistantly between the two values, which satisfy that a Bonferroni test has either 5% or 99% power to reject the global null hypothesis in a setting with no conservatism.

Cauchy Combination Test, Harmonic Mean p-value, Compound Bonferroni Arithmetic Mean and Bonferroni test all work well when signals are sparse, but have low power when there are many weak signals. In contrast, the Fisher Combination Test and Truncated Product Method work well when signals are dense and weak, but have almost no power when signals are sparse and strong. Thus, the $TMTI_\infty$ and $tTMTI_\infty$ appear to have high power against all alternative hypotheses, mimicking the properties of, e.g., the ACAT-O, a test which is shown to have high power against both sparse and dense alternatives (Liu et al., 2019). The ACAT-O, however, is designed specifically for sequencing studies and works by leveraging information about the minor-allele counts of a sequencing study, and thus cannot be directly applied in other settings. In contrast, $TMTI_\infty$ and $tTMTI_\infty$ work as regular combination tests and can be applied to any type of data, given that the assumption of independence is satisfied.

Second, $TMTI_1$ and $rtTMTI_\infty$ work well when signals are sparse, although not better than a Bonferroni correction. When signals are dense, these methods have less power than the $TMTI_\infty$, $tTMTI_\infty$, Fisher Combination Test and Truncated Product Method, but higher power than the Cauchy Combination Test and the Harmonic Mean p-value. The $TMTI_1$ and Rank Truncated Product Method have almost identical power

in all settings.

Third, all methods are, in some degree, affected by conservatism, in the sense that all methods generally have less power when the p-values from true hypotheses are conservative. When there are few false hypotheses ($N_{\text{false}} \leq 10^2$), the Fisher Combination Test and Truncated Product Method have almost no power even under mild conservatism. When there are sufficiently many false hypotheses ($N_{\text{false}} = 10^4$), the effect of conservatism is less pronounced. Overall, it appears that the TMTI tests are less affected by conservatism than the Fisher tests.

It is in line with the intuition behind the TMTI tests that TMTI_∞ and tTMTI_∞ do not perform as well as its competitors in situations where signals are sparse but strong, because these achieve their power from 'too many' of the marginal hypotheses being false. In contrast, minimum-p based tests, such as the Bonferroni procedure, need only a single, very strong signal to detect that the global null is false. Similarly, Liu and Xie (2020) argue that the Cauchy Combination Test only makes use of the first few small p-values to represent the overall significance. The same holds true for the Harmonic Mean p-value (Wilson, 2019). The reason that TMTI_1 and rtTMTI_∞ still perform similarly to these three methods is that the first local minimum of the sorted and transformed p-values is likely to lie early on when there are only a few, very small p-values, and likely to coincide with the global minimum of the K smallest p-values, if K is sufficiently small. Thus, TMTI_1 and rtTMTI_∞ share the property, that they are influenced the most by a few of the smallest p-values. The global minimum, however, need not lie early on, when there are only a few false hypotheses, meaning that the few signals that we do observe can potentially be missed when assessing the overall significance using TMTI_∞ or tTMTI_∞.

In B.3 we repeated the experiment of this section in a setting with non-constant μ values (i.e., when $N_{\text{false}} > 1$, the values of μ_{false} were allowed to differ for each false marginal hypothesis), finding results similar to those shown in Figure 2.

5 Multiple testing and strong FWER control

In this section, we consider a common task in statistics. Given p-values for a collection of hypotheses, which hypotheses can safely be rejected? As each p-value gives marginal Type I error control by definition, a naive approach would be to set a level, α, and reject any hypothesis if its corresponding p-value falls below α. However, as the number of tests conducted increases, more Type I errors will be made, which makes it necessary to employ methods that control for multiple testing. Popular targets one may wish to control for include the False Discovery Rate (Benjamini and Hochberg, 1995) and the Family-Wise Error Rate (FWER). To control the FWER the Bonferroni correction is often used, as it is easy to implement and guarantees strong FWER control. However, the Bonferroni correction has received criticism for, among other things, heavily increasing the risk of making Type II errors, i.e., failing to reject false hypotheses (Perneger, 1998). A general approach for turning global testing procedures into a procedure that controls the FWER for elementary hypotheses is the Closed Testing Procedure of Marcus et al. (1976). We briefly review the theory on Closed Testing Procedures.

Definition 2. *Let $\mathcal{I}$ be a set of indices and let $\{H_i\}_{i\in\mathcal{I}}$ denote a collection of hypotheses. For any subset $\mathcal{J} \subseteq I$, let $H_\mathcal{J} \coloneqq \bigcap_{j\in\mathcal{J}} H_j$ be the joint hypothesis. Let $\phi^\mathcal{J}$ be a random variable on $[0,1]$ satisfying*

$$H_\mathcal{J} \text{ true} \implies \forall \alpha \in [0,1]: \qquad \mathbb{P}(\phi^\mathcal{J} \leq \alpha) \leq \alpha.$$

That is, $\phi^\mathcal{J}$ is a valid p-value for the test of the joint hypothesis $H_\mathcal{J}$. Furthermore, for any subset $\mathcal{J} \subseteq \mathcal{I}$

we define the ***closure*** *of $\mathcal{J}$ in $\mathcal{I}$ to be*

$$\mathcal{J}^*_{\mathcal{I}} \coloneqq \bigcup_{\mathcal{K}:\mathcal{K}\subseteq\mathcal{I},\ \mathcal{J}\subseteq\mathcal{K}} \{\mathcal{K}\}.$$

That is, $\mathcal{J}^*_{\mathcal{I}}$ is the set of all supersets of $\mathcal{J}$ that are contained in $\mathcal{I}$. A ***Closed Testing Procedure*** *for the test of the joint hypothesis $H_{\mathcal{J}}$ is one that rejects $H_{\mathcal{J}}$ at level α if and only if every superset of $\mathcal{J}$ in $\mathcal{I}$ is also rejected at level α, i.e.,*

$$(H_{\mathcal{J}} \text{ is rejected at level } \alpha) = \bigcap_{\mathcal{K}\in\mathcal{J}^*_{\mathcal{I}}} \left(\phi^{\mathcal{K}} \leq \alpha\right).$$

That is, the event that we reject $H_{\mathcal{J}}$ occurs if and only if we reject all supersets of $\mathcal{J}$ in $\mathcal{I}$ marginally.

From the above, we see that a Closed Testing Procedure is more strict than marginal testing. That is, it becomes more difficult to reject any hypothesis, as we now need to reject all hypotheses that include the hypothesis of interest – not only the hypothesis itself. The upside is that we obtain strong control of the FWER.

Theorem 2 (Marcus et al. (1976))**.** *Let $\mathcal{J}_1, \ldots, \mathcal{J}_m$ be distinct subsets of a larger set of indices $\mathcal{I}$. Testing $H_{\mathcal{J}_1}$ through $H_{\mathcal{J}_m}$ each at level α by means of a closed testing procedure controls the FWER at level α in the strong sense.*

Given any general method to construct tests of joint hypotheses from marginal tests, we can employ these in a Closed Testing Procedure to obtain strict control of the FWER. It is generally accepted that Closed Testing Procedures are more powerful than other methods that control the FWER (Grechanovsky and Hochberg, 1999), although this power increase comes at the cost of a heavy computational burden. Given m marginal hypotheses which we want to test, we need to perform $\sum_{i=1}^{m}\binom{m}{i} = 2^m - 1$ tests. This is because we need to test all possible intersection hypotheses, which corresponds to the powerset of all hypotheses, minus the empty set. Thus, in many cases, it is not feasible to perform a Closed Testing Procedure when m is even slightly large. Indeed, even with just $m = 300$ marginal tests, the number of tests to be performed in a full Closed Testing Procedure is $2^{300} - 1 \approx 2 \cdot 10^{90}$ – roughly 10 billion times the number of atoms in the observable universe. Thus, with many procedures, one seeks to find a shortcut so that only a subset of the powerset of hypotheses needs to be tested. This is often possible (Grechanovsky and Hochberg, 1999) and considerably reduces the computational complexity of carrying out a Closed Testing Procedure.

Zaykin et al. (2002) introduce a shortcut for the Truncated Product Method, reducing the computational complexity of the Closed Testing Procedure from $\mathcal{O}(2^m)$ to $\mathcal{O}(m^2)$. In a recent result, Tian et al. (2021) give the same shortcut for a family of combination tests that are sums of marginal tests. Dobriban (2020) gives a shortcut for test statistics that are monotone and symmetric. Here, we provide a shortcut for class of combination tests, which are monotone but not necessarily symmetric, and not necessarily sums of marginal tests. Furthermore, we show that TMTI_∞, tTMTI_∞ and rtTMTI_∞ all admit this shortcut.

Lemma 2. *Let $\mathbf{p}_{\mathcal{I}}$ be a set of observed p-values with $\mathcal{I} \coloneqq \{1, \ldots, m\}$ and $m \geq 2$. Let $\mathcal{J}^k$ be the set of all subsets of $\mathcal{I}$ with $|\mathcal{J}| = k$. Let $\mathcal{X} \subseteq \mathbb{R}$ be a set and let $F_{(1)} : [0,1] \to \mathcal{X}, \ldots, F_{(k)} : [0,1] \to \mathcal{X}$ be a sequence of functions that satisfy*

$$\forall j \in \mathcal{I}\ \ \forall x \in \mathcal{X}\ \ \forall \epsilon \geq 0 : \qquad F_{(j)}(x) \leq F_{(j)}(x+\epsilon) \tag{C1}$$

and

$$\forall j \in \{1, \ldots, m-1\}\ \ \forall x \in \mathcal{X} : \qquad F_{(j)}(x) \geq F_{(j+1)}(x). \tag{C2}$$

Define for all $\mathcal{J} \in \mathcal{J}^k$ the random variable $\boldsymbol{Y}^{\mathcal{J}} \coloneqq (F_{(1)}(p_{(1)}), \dots, F_{(k)}(p_{(k)}))$ and let $h : \mathcal{X}^k \to [0,1]$ be a function satisfying

$$\forall \boldsymbol{x} \in \mathcal{X}^k \;\; \forall \boldsymbol{\epsilon} \in \mathbb{R}_+^k : \qquad h(\boldsymbol{x}) \leq h(\boldsymbol{x} + \boldsymbol{\epsilon}), \tag{C3}$$

Let $\eta : \mathcal{I} \to \mathcal{I}$ be a bijection ordering $\boldsymbol{p}_{\mathcal{I}}$, i.e., $p_{\eta(1)} \leq \dots \leq p_{\eta(m)}$. It then follows that for any two sets, $\mathcal{J}_1, \mathcal{J}_2 \in \mathcal{J}^k$

$$\eta(\mathcal{J}_1) \leq \eta(\mathcal{J}_2) \implies h(\boldsymbol{Y}^{\mathcal{J}_1}) \leq h(\boldsymbol{Y}^{\mathcal{J}_2}).$$

In the above, the operation $\leq$ applied to the sets $\eta(\mathcal{J}_1)$ and $\eta(\mathcal{J}_2)$ is taken to mean element-wise less than or equal to.

Lemma 2 states that whenever we consider k p-values, we will obtain a smaller test statistic if we substitute one or more of them with smaller p-values. In the context of closed testing, this implies that when considering the closure of an atom, say $\{j\}$, then among all subsets of size k in $\{j\}^*_{\mathcal{I}}$ we do not need to test all $m!/((m-k)!k!)$ intersection hypotheses. This is because we know that the largest (smallest) test statistic is obtained by considering the p-value p_j combined with the $k-1$ largest (smallest) remaining p-values. Assuming that the underlying distribution of the p-values is exchangeable, this implies that the p-values for the combination tests obey the same inequalities as the test statistics, and thus we need only consider the intersection hypothesis which we know will yield the largest p-value.

Remark 4. *The same result as in Lemma 2 can be obtained by reversing the inequalities in Equations* (C1)*,* (C2) *and* (C3)*. In contrast, we can obtain a version which gives*

$$\eta(\mathcal{J}_1) \leq \eta(\mathcal{J}_2) \implies h(\boldsymbol{Y}^{\mathcal{J}_1}) \geq h(\boldsymbol{Y}^{\mathcal{J}_2})$$

if we reverse the inequalities in Equations (C1) *and* (C2) *and keep Equation* (C3)*, or if we reverse the inequality in Equation* (C3) *and keep Equations* (C1) *and C2. The choice of which version to use depends on whether small or large values of the test statistic are critical.*

Theorem 3. *TMTI$_\infty$, tTMTI$_\infty$ and rtTMTI$_\infty$ all satisfy the conditions of Lemma 2.*

Remark 5. *Even though the TMTI$_\infty$ variants all satisfy the conditions in Lemma 2, not all TMTI$_n$ variants do. To see this, consider two sets of p-values, $\boldsymbol{p}_1 = (0.25, 0.50, 0.75)$ and $\boldsymbol{p}_2 = (0.2, 0.5, 0.75)$. Then $\boldsymbol{Y}_1 = (0.58, 0.5, 0.42)$, making 0.42 the first local minimum of $\boldsymbol{Y}_1$, and $\boldsymbol{Y}_2 = (0.49, 0.5, 0.42)$, making 0.49 the first local minimum of $\boldsymbol{Y}_2$. Thus, we have $\boldsymbol{Y}_2 \leq \boldsymbol{Y}_1$ but $h_{TMTI_1}(\boldsymbol{Y_2}) > h_{TMTI_1}(\boldsymbol{Y_1})$.*

Theorem 4. *Let $p_{\mathcal{I}}$, $F_{(1)}, \dots, F_{|\mathcal{I}|}$, $\mathcal{X}$, h and $\boldsymbol{Y}^{\mathcal{J}}$ be defined as in Lemma 2. Assume that the underlying distribution of $p_{\mathcal{I}}$ is exchangeable. If Equations* (C1)*,* (C2) *and* (C3) *are satisfied, then a Closed Testing Procedure using $h(\boldsymbol{Y}^{\mathcal{J}})$ as test statistic can be used to obtain control of the FWER for all marginal hypotheses in at most $\frac{1}{2}m(m-1)$ steps.*

Remark 6. *The result in Lemma 2 and its converse in Remark 4 does not only apply to TMTI statistics. For example, letting $F_{(1)}, \dots, F_{(m)}$ be the identity mappings and $h(\boldsymbol{x}) \coloneqq -2\sum_{i=1}^m \log x_i$ we obtain the Fisher Combination Test, which then gives us the well-known $\mathcal{O}(m^2)$ shortcut described e.g.,in Zaykin et al. (2002). Similarly, letting $h(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^m \tan((0.5 - x_i)\pi)$, we obtain the unweighted Cauchy Combination Test.*

In Algorithm 1, we give an example of how this shortcut procedure can be implemented to return adjusted p-values for the tests of all marginal hypotheses.

If the practitioner is content with having only a lower bound on the adjusted p-value, whenever the test is not rejected at a chosen level, Theorem 4 provides an upper bound on the number of steps required

Algorithm 1: Shortcut Closed Testing Procedure for statistics satisfying Conditions (C1), (C2) and (C3)

Input: Sorted p-values $p_1 \leq \cdots \leq p_m$ for tests of hypotheses $H_1, \ldots, H_m$, a significance level α
Output: Adjusted p-values for the tests of $H_1, \ldots, H_m$
Construct an empty $m \times m$ matrix Q.
for $i = 1, \ldots, m-1$ **do**
  $c \leftarrow m$.
  $Q_{i,c} \leftarrow p_i$.
  for $j = m, \ldots, i+1$ **do**
    $c \leftarrow c - 1$.
    Test the hypothesis $\left(\bigcap_{k=m}^{j} H_k\right) \cap H_i$ and save the p-value as $p_{i,m:j}$.
    $Q_{i,c} \leftarrow p_{i,m:j}$.
for $i = 1, \ldots, m-1$ **do**
  $Q_{i,(i+1):m} \leftarrow Q_{i,i}$
  $\tilde{p}_i \leftarrow \max Q_{1:m,i}$
$\tilde{p}_m \leftarrow \max Q_{1:m,m}$
Return $\tilde{p}_1, \ldots, \tilde{p}_m$ as adjusted p-values for $H_1, \ldots, H_m$.

to complete the Closed Testing Procedure. For instance, if the global hypothesis cannot be rejected, no marginal hypothesis can be rejected, and the procedure can therefore be stopped and the p-value for the test of the global hypothesis can be used as a lower bound for all adjusted p-values. Stopping the procedure early can speed up computations considerably, especially when m is large but very few hypotheses can be rejected. Furthermore, Lemma 2 also implies that an adjusted p-value for a single elementary hypothesis can be obtained in only m steps. In practice, one is often only interested in obtaining adjusted p-values for the hypotheses, for which the marginal p-value is significant (as the remaining hypotheses cannot be rejected in a Closed Testing Procedure). Thus, if there are, say, n elementary hypotheses for which the marginal p-value is significant, these can be adjusted in $\mathcal{O}(nm)$ complexity.

We have exemplified in Figure 3 how the reduced test-tree looks for the Closed Testing Procedure when applying any test that satisfies the conditions of Lemma 2 in the case of $m = 4$ total tests, where the hypotheses are labeled such that H_i corresponds to the ith lowest p-value. To obtain an adjusted p-value for any marginal hypothesis, say H_i, one takes the maximal p-value from the test of all ancestral hypotheses in the graph. For example, the adjusted p-value for the test of H_2 would be the largest of the p-values from the tests of H_2, $H_{2,4}$, $H_{2,3,4}$ and $H_{1,2,3,4}$.

5.1 A remark on mixture strategies in Closed Testing Procedures

Definition 2 of a Closed Testing Procedure and the subsequent Theorem 2 on FWER control make no assumptions on the choice of local tests, $\phi^{\mathcal{J}}$, and these can in principle vary across all intersection hypotheses to be tested. We only require that every local test, $\phi^{\mathcal{J}}$, is a valid α-level test. The natural choice is to use the same kind of test at every intersection hypothesis, e.g., TMTI$_\infty$. However, we can in principle employ any choice of local test. In some cases, we argue, it is reasonable to use different local tests for different kinds of intersection hypotheses. When we go through a Closed Testing Procedure, we are going to consider tests of many different kinds of hypotheses, and in particular different kinds of alternative hypotheses. As previously discussed, TMTI$_\infty$ has slightly lower power than other methods in situations where signals are

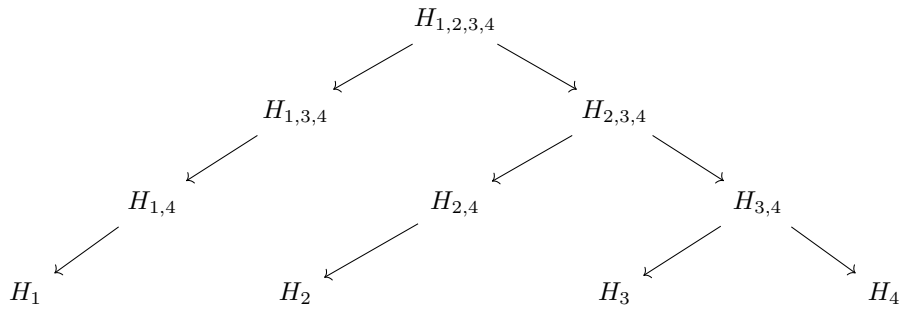


Figure 3: Test procedure with $m = 4$

extremely sparse. According to the shortcut strategy outlined in Algorithm 1, we need only consider the $|\mathcal{J}| - 1$ largest p-values alongside the jth p-value, when testing all supersets of $\{j\}$ of size $|\mathcal{J}|$. When $|\mathcal{J}|$ is small or when false hypotheses are sparse, it is likely that these intersection hypotheses will consist of a single false hypothesis (if any) with all the remaining hypotheses being true. It makes sense, then, to employ a different test in these situations, which has power against alternatives of sparse signals. However, this alternative test needs to satisfy the same shortcut as TMTI_∞ in order to be employed across all supersets of equal size. One such choice is rtTMTI_∞ with a low choice of K – e.g., $K = 1$ – as this method satisfies the same shortcut as TMTI_∞, but has higher power against sparse alternatives, as discussed in Section 4. Given a number of hypotheses, say m, of which we expect $F < m$ to be false, we could for example conduct the Closed Testing Procedure by employing rtTMTI_∞ with a small K whenever we consider supersets of size at most $m - F$ and TMTI_∞ whenever we consider supersets of size greater than $m - F$. We call such a strategy a mixture Closed Testing Procedure. We return to mixture Closed Testing Procedures in Section 8, where we analyse a real dataset. The reasoning is that once we start considering supersets of size greater than $m - F$, the intersections hypotheses considered in the shortcut procedure will potentially include multiple false hypotheses, while they will include at most one false hypothesis when considering supersets of size less than $m - F$.

We stress that when employing a mixture Closed Testing Procedure, the choice of local tests should be made *a priori* and not be data-driven, so as not to incur new multiplicity problems.

6 The number of false hypotheses in a rejection set and k-FWER control

Given a set of hypotheses, indexed by $\mathcal{J}$, such that we can safely reject the joint hypothesis $H_\mathcal{J}$ – i.e., we conclude that at least one of the hypotheses in $\mathcal{J}$ is false – the natural question is then how many of the hypotheses in $\mathcal{J}$ are false. To answer this, Goeman and Solari (2011) provide a simple way of generating a $1 - \alpha$ confidence set for the number of false hypotheses contained in $\mathcal{J}$ when using Closed Testing Procedures. Let $\mathcal{R} \subseteq \text{powerset}(\mathcal{J})\backslash\{\emptyset\}$ be the set of all intersection hypotheses that can be rejected by any Closed Testing Procedure. Define τ to be the number of true hypotheses in $\mathcal{J}$ and $t_\alpha := \max\{|\mathcal{K}| : \mathcal{K} \subseteq \mathcal{J}, \mathcal{K} \notin \mathcal{R}\}$ to be the size of the largest intersection hypothesis in $\mathcal{J}$ that can not be rejected by the Closed Testing Procedure.

Theorem 5 (Goeman and Solari (2011))**.** *The sets* $\{0, \ldots, t_\alpha\}$ *and* $\{|\mathcal{J}| - t_\alpha, \ldots, |\mathcal{J}|\}$ *are* $1-\alpha$ *confidence sets for the number of true hypotheses,* τ*, and the number of false hypotheses,* $|\mathcal{J}| - \tau$*, respectively.*

This remarkably simple theorem has the implication, that we can generate confidence sets for the number of false hypotheses among all those tested in only t_α steps when using any test procedure satisfying the conditions of Lemma 2, assuming that the p-values are realized from an exchangeable distribution. We describe in Algorithm 2 how to do this.

The quantity t_α depends on the choice of test used on each intersection. Different tests have power against different alternatives, and a test that has low power for a particular intersection hypothesis will be more likely to not reject that hypothesis. Thus, if the chosen test has low power for the particular data, the resulting confidence set for the number of false hypotheses will be conservative. In contrast, if the chosen test has high power against the particular alternative, the confidence set tightens.

As noted in Goeman and Solari (2011), we can also apply Theorem 5 as a way of controlling the k-FWER, i.e., the probability of making at least k false rejections. To control the k-FWER, we find the largest i such that $t_\alpha < k$ when calculated for the set of hypotheses yielding the i smallest p-values. That is, if we find that F hypotheses in a set, say $\mathcal{J}$, are false with $1-\alpha$ confidence, then we can reject every hypothesis in that set while controlling the k-FWER at $k = |\mathcal{J}| - F + 1$. This can be done in $\mathcal{O}(m^3)$ time. This is because determining t_α from a set $\mathcal{J}$ is done in $\mathcal{O}(|\mathcal{J}|^2)$ time, as described in Algorithm 2, and we now need to do this for subsets of increasing (or decreasing, depending on the search direction) size. It is worth noting, that the k at which the practitioner wishes to control the k-FWER need not be chosen *a priori*, as the confidence sets are simultaneous for all choices of $\mathcal{J}$, simply because the closure of each $\mathcal{J}$ is contained within the full closure.

7 Additional computational considerations

The computational time of a carrying out a Closed Testing Procedure when using the above shortcuts is manageable for reasonable values of m. For example, computing adjusted p-values for a set of $m = 100$ p-values take roughly two seconds on a standard laptop using single-threaded computations. Still, there is a considerable amount of computational effort involved in carrying out a Closed Testing Procedure when m is sufficiently large, in part because the CDFs of the TMTI statistics will have to be bootstrapped. To further reduce the computational burden, we offer the following result.

Lemma 3. *Let* $\mathcal{J}_1$ *and* $\mathcal{J}_2$ *be sets such that* $\mathcal{J}_1 \subsetneq \mathcal{J}_2$*. Then*

$$\forall x \in (0,1): \qquad \gamma^{\mathcal{J}_1}(x) < \gamma^{\mathcal{J}_2}(x).$$

That is, a conservative p-value for the test of $H_{\mathcal{J}_1}$ *can be obtained by using* $\gamma^{\mathcal{J}_2}$ *instead of* $\gamma^{\mathcal{J}_1}$ *when computing the p-value.*

The purpose of Lemma 3 is that the user can choose to skip the bootstrap at several layers of the Closed Testing Procedure and instead simply use the CDF of a higher layer. This improves the running time of the algorithm at the cost of using conservative p-values at the layers where the bootstrap was skipped. Exactly how costly this trade-off is, depends on how many layers are skipped each time. We conjecture that the p-value will only be slightly conservative if the number of layers skipped is small relative to the size of the subsets considered.

Algorithm 2: Confidence set for the number of false hypotheses among $\mathcal{J} \subseteq \mathcal{I}$

Input: Hypotheses $(H_i)_{i\in\mathcal{I}}$, an ordered set $\mathcal{J} \subseteq \mathcal{I}$, ordered p-values $p_1 \leq \cdots \leq p_{|\mathcal{I}|}$.
Output: A $1-\alpha$ confidence set for the number of false hypotheses in $\mathcal{J}$.
Initialization
  $t_\alpha \leftarrow 0$
  $m \leftarrow |\mathcal{J}|$
  $\tilde{p} \leftarrow (p_j)_{j\in\mathcal{J}}$
if $m = |\mathcal{I}|$ **then**
  for $i = m, \ldots, 1$ **do**
    $\tilde{\mathcal{J}} \leftarrow \{m-i+1, \ldots, m\}$
    Let $p_{\tilde{\mathcal{J}}}$ be the p-value for the test of $\bigcap_{j\in\tilde{\mathcal{J}}} H_j$.
    if $p_{\tilde{\mathcal{J}}} \geq \alpha$ **then**
      Set $t_\alpha \leftarrow i$
      Break the loop and return $\{|\mathcal{J}| - t_\alpha, \ldots, |\mathcal{J}|\}$.
else
  for $i = m, \ldots, 1$ **do**
    $\tilde{\mathcal{J}} \leftarrow \{m-i+1, \ldots, m\}$
    $\hat{p} \leftarrow (p_i)_{i\in\mathcal{I}\setminus\tilde{\mathcal{J}}}$
    Let $p_{\tilde{\mathcal{J}}}$ be the p-value for the test of $\bigcap_{j\in\tilde{\mathcal{J}}} H_j$
    for $j = 1, \ldots, |\mathcal{I}\setminus\tilde{\mathcal{J}}|$ **do**
      Define $\tilde{\mathcal{J}}_2$ as $\tilde{\mathcal{J}}$ appended with the j largest values of $\hat{\mathcal{I}}$.
      Update $p_{\tilde{\mathcal{J}}}$ as the p-value from the test of $\bigcap_{j\in\tilde{\mathcal{J}}_2} H_j$.
      if $p_{\tilde{\mathcal{J}}} \geq \alpha$ **then**
        Set $t_\alpha \leftarrow i$
        Break the loop and return $\{|\mathcal{J}| - t_\alpha, \ldots, |\mathcal{J}|\}$.
Return $\{|\mathcal{J}| - t_\alpha, \ldots, |\mathcal{J}|\}$.

8 An example using real data

In this section, we give an example of how TMTI$_\infty$ performs against other methods when applied to real data. For this purpose, we consider data from the *National Assessment of Educational Progress* on the state-wise changes in eighth-grade mathematics achievements from 1990 to 1992. This data is, among other places, presented in Williams et al. (1999), where the authors compute two-sided T-tests for the mean change in mathematical achievements over the two-year period, to quantify whether or not any particular state has progressed or worsened during that time period. The original data includes one p-value of exactly 0. We have rounded this to be 0.00001 here to ensure Type I error control. The same data is used in Benjamini and Hochberg (2000), where the authors find that mathematics achievements have changed significantly in 24 of the 34 states. However, the authors control the more lenient False Discovery Rate, and thus their results are not directly comparable to the ones presented here. Williams et al. (1999) apply a significance level of 0.025 instead of the usual 0.05. In Benjamini and Hochberg (2000), the authors apply the usual 0.05 significance level but have doubled all p-values such that the results are comparable. We have done the same here. The data, adapted from Williams et al. (1999), is presented in Table 1. Given that we only have access to summary statistics, we assume that all p-values are independent. Whether this is a reasonable assumption can be debated.

There are several questions regarding this data that may be of interest to the practitioner:

1. Did mathematics achievements change significantly in any state from 1990 to 1992?
2. In how many states did mathematics achievements change significantly?
3. In what states did mathematics achievements change significantly?

To answer the first question, we can, for example, apply TMTI$_\infty$ to obtain a p-value. Doing so results in a p-value of 1.58×10^{-13}. Thus, we find evidence that mathematics education has changed significantly in at least one of the 34 states.

To answer the remaining two questions, we apply TMTI$_\infty$ in a Closed Testing Procedure as well as a mixture Closed Testing Procedure, using rtTMTI$_\infty$ with $K = 1$ whenever we consider fewer than 15 hypotheses and TMTI$_\infty$ when considering more. For comparison, we also apply the standard Bonferroni correction as well as the Fisher Combination Test in a Closed Testing Procedure and a Rank Truncated Product Method/Fisher Combination Test (denoted rTPM/FCT) mixture, using the Rank Truncated Product Method with $K = 1$ for intersection hypotheses of size at most 15. The choice of $K = 1$ for intersection hypotheses smaller than 15 corresponds to a belief that at most 15 hypotheses are true. Here, we have chosen 15 at random, but a practitioner with subject matter knowledge can choose this based on prior knowledge.

By applying Algorithm 2 with TMTI$_\infty$ we find that $\{23, \dots, 34\}$ is a 95% confidence set for the number of states in which mathematics achievements have changed. In contrast, using the Fisher Combination Test, the confidence set is $\{19, \dots, 34\}$. The same confidence set is found when applying the mixture strategies. That is, we can say with 95% confidence, that mathematics achievements have changed significantly in at least 23 states when using TMTI$_\infty$. The improved performance of TMTI$_\infty$ over the Fisher Combination Test here is likely due to TMTI$_\infty$ generally having high power across a wide range of settings (see Section 4), whereas the Fisher Combination Test lacks power in settings with sparse, strong signals. When carrying out a Closed Testing Procedure, we test many different kinds of joint hypotheses (i.e., some containing more false hypotheses than others), and it is thus beneficial that the employed test has high power in many different settings.

To determine which of the hypotheses we can say with certainty are false while controlling the FWER, we applied Algorithm 1. Here, TMTI$_\infty$, the Fisher Combination Test and the Bonferroni correction perform

identically and are capable of rejecting the bottom four hypotheses. In contrast, when we apply the two mixture strategies, we can reject the bottom seven hypotheses. Thus, by incorporating prior knowledge, we can increase the size of the rejection set for which we have FWER control.

To find which of the hypotheses can be rejected while controlling the more general k-FWER, we consider sets of increasing size of the smallest p-values, each time calculating t_α. For any chosen set, say $\mathcal{J}$, we can reject the entire set while controlling the k-FWER at $t_\alpha + 1$. Doing so, we find that the 11 hypotheses giving rise to the smallest p-values can be rejected by the TMTI_∞ while controlling the k-FWER at $k = 2$. In contrast, the mixture Closed Testing Procedure is only capable of rejecting the bottom eight hypotheses while controlling the k-FWER at $k = 2$. If we are willing to accept a more lenient k-FWER control of $k = 5$, we are capable of rejecting the bottom 22 hypotheses using the TMTI_∞ and the bottom 11 hypotheses using the $\text{rtTMTI}_\infty/\text{TMTI}_\infty$ mixture. In Figure 4, we have summarized the associated k at which we control the k-FWER at, when rejecting the bottom t hypotheses, for $t = 1, \ldots, 34$, when using TMTI_∞, the $\text{rtTMTI}_\infty/\text{TMTI}_\infty$ mixture, the Fisher Combination Test and the rTPM/FCT mixture, respectively. Here, we note that TMTI_∞ is weakly better than the Fisher Combination Test except for at a single rejection set, $\{1, \ldots, 13\}$.

State	GA	AR	AL	NJ	NE	ND	DE	MI
p-value (%)	85.628	60.282	44.008	41.998	38.640	36.890	31.162	23.522
	LA	IN	WI	VA	WV	MD	CA	OH
	20.964	19.388	15.872	14.374	10.026	8.226	7.912	6.590
	NY	PA	FL	WY	NM	CT	OK	KY
	5.802	5.572	5.490	4.678	4.650	4.104	2.036	0.964
	AZ	ID	TX	CO	IA	NH	NC	HI
	0.904	0.748	0.404	0.282	0.200	0.180	0.002	0.002
	MN	RI						
	0.002	0.001						

Table 1: States and their p-values for T-tests of changes in mathematics achievements from 1990 to 1992.

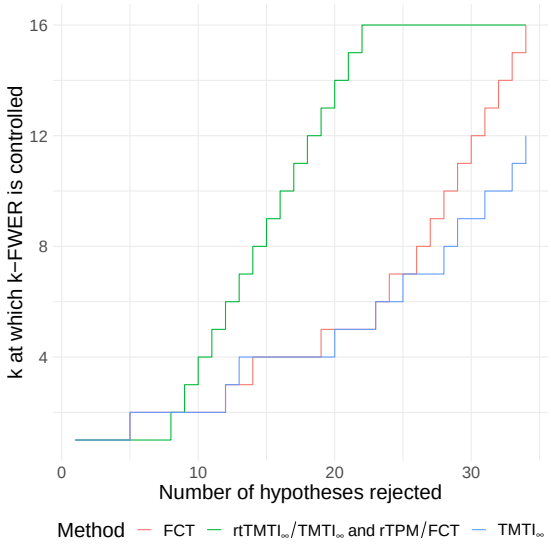


Figure 4: Overview of the different Closed Testing Procedure methods employed and the k at which they control the k-FWER, when rejecting the bottom t hypotheses, for $t = 1, \ldots, 34$. The two mixture strategies, rtTMTI$_\infty$/TMTI$_\infty$ and rTPM/FCT, are colored the same, as their results are identical.

That the two mixture strategies have higher power to detect differences when controlling the FWER than TMTI$_\infty$ and the Fisher Combination Test, but lower power to detect differences when using the more lenient k-FWER control, may seem counter-intuitive and requires some exposition. The difference lies in what intersection hypotheses need to be considered in the full Closed Testing Procedure test tree. When controlling the FWER, we are in principle looking through all intersection hypotheses, and then we use the maximal p-values along the closure of each atom as the adjusted p-values. As outlined in Lemma 2, however, we need only consider the part of the closure that contains subsets containing only the atom and the largest p-values. When we apply Algorithm 2 iteratively to obtain k-FWER control, we are considering the closures, not of atoms, but of intersections. Put differently, consider the index set $\mathcal{J} := \{1, \ldots, t\}$, for some t, of the t smallest p-values. We are then going to consider the closure of $\{t\}$ in $\mathcal{J}$, i.e., $\{t\}^*_\mathcal{J}$ at first. For all sets in this closure, say $\tilde{\mathcal{J}} \in \{t\}^*_\mathcal{J}$, we are then going to calculate the adjusted p-value as the maximal p-value along the closure of $\tilde{\mathcal{J}}$ in $\mathcal{I}$, i.e., $\tilde{\mathcal{J}}^*_\mathcal{I}$. These sets are not, as they were in the ordinary Closed Testing Procedure, sets consisting of an atom unioned with the largest p-values, but rather several, possibly neighboring, atoms unioned with the largest p-values. We constructed the mixture strategies to have higher power in situations in which we considered intersection hypotheses with only a single false hypothesis in them. Using this method, we are now considering sets that possibly have multiple false hypotheses in them, even when the total number of hypotheses included in the set is low – which is when TMTI$_\infty$ gains its power. In contrast, rtTMTI$_\infty$ with a small K loses power, when there are more than K false hypotheses present in the intersection hypothesis.

A detailed table with adjusted p-values for all of the tests employed here can be found in C.

9 Conclusion

We have introduced the 'Too Many, Too Improbable' (TMTI) family of combination test statistics for testing joint hypotheses among m marginal hypotheses. The TMTI family includes truncation-based tests, similar to those of Zaykin et al. (2002) and Dudbridge and Koeleman (2003), for testing global hypotheses against sparse alternatives. We have shown in Section 4 that the TMTI tests outperforms other combination tests in many situations. In particular, we found that TMTI_∞ and tTMTI_∞ were the only tests that were able to achieve high power both when signals are dense but weak and when signals are sparse but strong. Although we found in all scenarios that there was at least one other test that performed equally as well as TMTI_∞ tTMTI_∞, no other combination tests had similar performance across all scenarios. This property is useful, e.g., if one has no *a priori* knowledge about the sparsity and strength of signals and for generating $1 - \alpha$ confidence sets for the number of false hypotheses in a rejection set.

In Section 5, we have given an $\mathcal{O}(m^2)$ shortcut for controlling the Family-Wise Error Rate using Closed Testing Procedures (Marcus et al., 1976) for a large class of test statistics, which includes the TMTI family of test statistics, but also the Cauchy Combination Test among others. Using this shortcut, we use the work of Goeman and Solari (2011) in Section 6 to develop an $\mathcal{O}(m^3)$ algorithm for controlling the generalized FWER as well as an $\mathcal{O}(m)$ algorithm for obtaining $1 - \alpha$ confidence sets for the number of false hypotheses among all hypotheses.

In Section 8, we applied a TMTI test in a Closed Testing Procedure, as well as a mixture – i.e., varying local tests across the Closed Testing Procedure – of two TMTI tests, to a real dataset and compared it to the Fisher Combination Test applied in a Closed Testing Procedure. Here we found that all TMTI tests were able to reject the same hypotheses as the Fisher Combination Test, but that the TMTI test generated a narrower confidence set for the number of false hypotheses among the collection of considered hypotheses. Additionally, we found that by employing mixture strategies, we were able to reject more hypotheses than with standard methods. However, the mixture strategies performed worse when controlling the k-FWER with $k \geq 2$.

Supplementary material

Proofs of all lemmas and theorems are supplied in the appendix. Furthermore, the appendix includes further simulations to support those of Section 4 and a detailed table containing the adjusted p-values of the procedures applied in Section 8.

The TMTI family of test statistics and the shortcuts for k-FWER and confidence sets is implemented in the R package `TMTI` available on CRAN.

References

Benjamini, Y., 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, 289–300.

Benjamini, Y., Hochberg, Y., 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of educational and Behavioral Statistics 25, 60–83.

Brown, M.B., 1975. 400: A method for combining non-independent, one-sided tests of significance. Biometrics , 987–992.

Chen, Y., Liu, P., Tan, K.S., Wang, R., 2020. Trade-off between validity and efficiency of merging p-values under arbitrary dependence. arXiv preprint arXiv:2007.12366 .

Dobriban, E., 2020. Fast closed testing for exchangeable local tests. Biometrika 107, 761–768.

Dudbridge, F., Koeleman, B.P., 2003. Rank truncated product of p-values, with application to genomewide association scans. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 25, 360–366.

Fisher, R.A., 1992. Statistical methods for research workers, in: Breakthroughs in statistics. Springer, pp. 66–70.

Goeman, J.J., Solari, A., 2011. Multiple testing for exploratory research. Statistical Science 26, 584–597.

Grechanovsky, E., Hochberg, Y., 1999. Closed procedures are better and often admit a shortcut. Journal of Statistical Planning and Inference 76, 79–91.

Kessy, A., Lewin, A., Strimmer, K., 2018. Optimal whitening and decorrelation. The American Statistician 72, 309–314.

Kost, J.T., McDermott, M.P., 2002. Combining dependent p-values. Statistics & Probability Letters 60, 183–190.

Littell, R.C., Folks, J.L., 1973. Asymptotic optimality of fisher's method of combining independent tests ii. Journal of the American Statistical Association 68, 193–194.

Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E., Lin, X., 2019. Acat: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. The American Journal of Human Genetics 104, 410–421.

Liu, Y., Xie, J., 2020. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. Journal of the American Statistical Association 115, 393–402.

Marcus, R., Eric, P., Gabriel, K.R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63, 655–660.

Perneger, T.V., 1998. What's wrong with bonferroni adjustments. Bmj 316, 1236–1238.

Peters, J., Bühlmann, P., Meinshausen, N., 2016. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78, 947–1012.

Šidák, Z., 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62, 626–633.

Tian, J., Chen, X., Katsevich, E., Goeman, J., Ramdas, A., 2021. Large-scale simultaneous inference under dependence. arXiv preprint arXiv:2102.11253 .

Vovk, V., Wang, R., 2020. Combining p-values via averaging. Biometrika 107, 791–808.

Williams, V.S., Jones, L.V., Tukey, J.W., 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. Journal of educational and behavioral statistics 24, 42–69.

Wilson, D.J., 2019. The harmonic mean p-value for combining dependent tests. Proceedings of the National Academy of Sciences 116, 1195–1200.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X., 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics 89, 82–93.

Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H., Weir, B.S., 2002. Truncated product method for combining p-values. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 22, 170–185.

A Proofs

A.1 Proof of Lemma 1

This holds trivially, as the first Y smaller than the following n must necessarily be larger when we only consider the first $c < |\mathcal{I}|$ values of $\boldsymbol{Y}^{\mathcal{I}}$ than if we consider the full sequence.

A.2 Proof of Theorem 1

Assume first that $c = m$. The order statistics $(P_{(1)}, \ldots, P_{(m)})$ have a constant joint density $m!$ on the simplex $\{\boldsymbol{p} \in [0,1]^m : p_1 \leq \cdots \leq p_m\}$. Thus

$$\begin{aligned}
\gamma^{\mathcal{I}}(x) \coloneqq& \mathbb{P}(\min(\beta(1,m)(P_{(1)}), \ldots, \beta(m,1)(P_{(m)})) \leq x) \\
=& 1 - \mathbb{P}(\beta(1,m)(P_{(1)}) > x, \ldots, \beta(m,1)(P_{(m)}) > x) \\
=& 1 - \mathbb{P}(P_{(1)} > \beta^{-1}(1,m)(x), \ldots, P_{(m)} > \beta^{-1}(m,1)(x)) \\
=& 1 - \mathbb{P}(P_{(1)} > x_1, \ldots, P_{(m)} > x_m) \\
=& 1 - m! \int_{x_m}^{1} \int_{x_{m-1}}^{q_m} \cdots \int_{x_2}^{q_3} \int_{x_1}^{q_2} dq_1 \, dq_2 \ldots dq_{m-1} \, dq_m. \qquad (5)
\end{aligned}$$

The integral in Equation (5) can be expressed as

$$\int_{x_m}^{1} \int_{x_{m-1}}^{q_m} \cdots \int_{x_2}^{q_3} \int_{x_1}^{q_2} dq_1 \, dq_2 \ldots dq_{m-1} \, dq_m = Q_m(1; (1, -\bar{Q}_{1,c}, \ldots, -\bar{Q}_{i-1,c})) - \bar{Q}_{m,c},$$

which in turn implies that

$$\begin{aligned}
\gamma^{\mathcal{I}}(x) &= 1 - m!(Q_m(1; (1, -\bar{Q}_{1,c}, \ldots, -\bar{Q}_{i-1,c})) - \bar{Q}_{m,c}) \\
&= 1 - m! \left(\frac{1}{m!}(1 - x_m^m) - \sum_{i=1}^{m-1} \bar{Q}_{i,c} \frac{1}{(m-i)!}(1 - x_m^{m-i}) \right) \\
&= x_m^m + \sum_{i=1}^{m-1} \bar{Q}_{i,c} \frac{m!}{(m-i)!}(1 - x_m^{m-i}).
\end{aligned}$$

Now, assume that $c < m$. Then the expression is identical to the one given in Equation (5), with the exception that the lower bound on all integrals after the c'th from the inside out will be x_c. The CDF is therefore

$$\gamma_{\infty,c}(x) = x_c^m + \sum_{i=1}^{m-1} \bar{Q}_{i,c} \frac{m!}{(m-i)!}(1 - x_c^{m-i}).$$

Now, let c be a random variable given by $c = \max\{i \in \{1, \ldots, m\} : p_{(i)} < \tau\}$. Assume without loss of generality that $p_1 \leq \cdots \leq p_m$. For any fixed integer $i \leq m$ we note that $p_1, \ldots, p_i \mid c = i \sim U(0, \tau)$ and that the joint distribution of $(p_{(1)}, \ldots, p_{(i)})$ conditional on $c = i$ has density $i!/\tau^i$ on the simplex $\{\boldsymbol{p} \in [0,1]^i : p_1 \leq \cdots \leq p_i < \tau\}$. By the Law of Total Probability, we can write

$$\gamma_{\infty,c}(x) = \sum_{i=0}^{m} \mathbb{P}(Z_{\infty,c} < x \mid c = i) \mathbb{P}(c = i).$$

The distribution of c is binomial with probability parameter τ and size m, i.e.,

$$\mathbb{P}(c = i) = \binom{m}{i} \tau^i (1-\tau)^{m-i}.$$

Consider first the case of $i \geq 1$. We see that

$$\begin{aligned}
\mathbb{P}(Z_{\infty,i} \leq x \mid c = i) &= 1 - \mathbb{P}(P_1 > x_1, \ldots, P_i > x_i \mid c = i) \\
&= 1 - \frac{i!}{\tau^i} \int_{x_i}^{\tau} \int_{x_{i-1}}^{q_i} \cdots \int_{x_2}^{q_3} \int_{x_1}^{q_2} dq_1 \, dq_2 \ldots dq_{i-1} \, dq_i I(x_i < \tau) \\
&= 1 - \frac{i!}{\tau^i} \left(\tilde{Q}_i - \bar{Q}_{i,m} \right) I(x_i < \tau).
\end{aligned}$$

In the case of $i = 0$, we see that $\mathbb{P}(Z_{\infty,c} < x \mid c = 0) = \mathbb{P}(\beta(1,m)(P_1) < x \mid c = 0)$. The distribution of P_1 conditionally on no p-values falling below τ is uniform on the interval $(\tau, 1)$. Thus, we have $\beta(1,m)(P_1) \mid c = 0 \sim U(\beta(1,m)(\tau), 1)$ and therefore

$$\mathbb{P}(Z_{\infty,c} < x \mid c = 0) = \frac{x - \beta(1,m)(\tau)}{1 - \beta(1,m)(\tau)} I(x_1 > x).$$

Combining all of the above, we obtain

$$\begin{aligned}
\gamma_{\infty,c}(x) = (1-\tau)^m &\frac{x - \beta(1,m)(\tau)}{1 - \beta(1,m)(\tau)} I(x_1 > x) \\
&+ \sum_{i=1}^{m} \left[\binom{m}{i} \tau^i (1-\tau)^{m-i} \left\{ 1 - \frac{i!}{\tau^i} \left(\tilde{Q}_i - \bar{Q}_{i,m} \right) I(x_i \leq \tau) \right\} \right]
\end{aligned}$$

which proves the claim.

A.3 Proof of Lemma 2

Assume without loss of generality that the p-values $\boldsymbol{p}_{\mathcal{I}}$ are already sorted, i.e., that $p_1 \leq \cdots \leq p_m$. Then η is the identity function and can therefore be omitted entirely. Fix a set $\mathcal{J} \coloneqq \{j_1, \ldots, j_k\} \in \mathcal{J}^k$ and assume without loss of generality that $j_1 < \cdots < j_k$. Fix $j \in \mathcal{J}$ and $l \in \mathcal{I} \backslash \mathcal{J}$ such that $l < j$ and let $\mathcal{J}_l^{-j} \coloneqq (\mathcal{J} \backslash \{j\}) \cup \{l\}$, i.e., $\mathcal{J}_l^{-j}$ is the set obtained by substituting j for l in $\mathcal{J}$. It suffices to show that

$$h(\boldsymbol{Y}^{\mathcal{J}_l^{-j}}) \leq h(\boldsymbol{Y}^{\mathcal{J}}).$$

There are two cases to consider:

Case 1; $j = \min \mathcal{J}$: If j is the smallest element in $\mathcal{J}$ then substituting it for l does not change the ordering of $\mathcal{J}$. By Condition (C1), it holds that

$$F_{(1)}(p_{(l)}) =: Y_1^{\mathcal{J}_l^{-j}} \leq Y_1^{\mathcal{J}} \coloneqq F_{(1)}(p_{(j)}).$$

As all other values for $\boldsymbol{Y}^{\mathcal{J}_l^{-j}}$ and $\boldsymbol{Y}^{\mathcal{J}}$ are unchanged, it follows from Condition C3 that $h(\boldsymbol{Y}^{\mathcal{J}_l^{-j}}) \leq h(\boldsymbol{Y}^{\mathcal{J}})$.

Case 2; $j > \min \mathcal{J}$**:** Define by $\tilde{j}$ the largest index in $\mathcal{J}$ smaller than j, i.e.

$$\tilde{j} \coloneqq \max\{i \in \mathcal{J} : i < j\}.$$

Suppose first that $\tilde{j} < l < j$. In this case, the ordering of $\mathcal{J}$ is unchanged when substituting j for l, making this case isomorphic to case 1. If, on the other hand, $l < \tilde{j}$ the ordering of $\mathcal{J}$ changes when substituting j for l. Let $\hat{j}$ be the smallest index in $\mathcal{J}$ larger than l, i.e.,

$$\hat{j} \coloneqq \min\{i \in \mathcal{J} : i > l\}.$$

Then we must show two things

1. $Y_l^{\mathcal{J}_l^{-j}} \leq Y_{\hat{j}}^{\mathcal{J}}$.
2. For all $i > \hat{j}$ it holds that $Y_i^{\mathcal{J}_l^{-j}} \leq Y_i^{\mathcal{J}}$.

Let $\eta_{\mathcal{A}} : \mathcal{A} \to \mathcal{A}$ be a function sorting the elements of a set $\mathcal{A}$. That is, $\eta_{\mathcal{A}}(a) = b$ if and only if a is the b'th lowest element in $\mathcal{A}$. We then see that $\eta_{\mathcal{J}_l^{-j}}(l) = \eta_{\mathcal{J}}(\hat{j})$ and thus that the first point above is satisfied as we have $p_{(l)} \leq p_{(\hat{j})}$ and therefore $Y_l^{\mathcal{J}_l^{-j}} \leq Y_{\hat{j}}^{\mathcal{J}}$ by Condition (C1). The second point above is satisfied by Condition (C2), as $\eta_{\mathcal{J}_l^{-j}}(h) = \eta_{\mathcal{J}}(h) + 1$ for any $h > \hat{j}$. This proves Lemma 2.

A.4 Proof of Theorem 3

We start by reminding the reader that all three TMTI statistics have $F_{(i)}(x) = \beta(i, m+1-i)(x)$, for $i = 1, \ldots, m$, regardless of the choice of n. These functions are weakly increasing, as they are CDFs, thus satisfying Condition (C1). Next, fix $x \in (0,1)$ and $i \in \mathcal{I}$ with $i < m$. We then see that

$$\begin{aligned}
F_{(i+1)}(x) &= \sum_{h=i+1}^{k} \binom{k}{h} x^h (1-x)^{k-h} \\
&= \beta(i+1, k+1-(i+1))(x) \\
&< \binom{k}{i} x^i (1-x)^{k-i} + \sum_{h=i+1}^{k} \binom{k}{h} x^h (1-x)^{k-h} \\
&= \sum_{h=i}^{k} \binom{k}{h} x^h (1-x)^{k-h} \\
&= \beta(i, k+1-i)(x) \\
&= F_{(i)}(x),
\end{aligned}$$

by using Equation (2). Thus, Condition (C2) is satisfied.

Let $\mu > 1$ and note that

$$\begin{aligned}
h_{TMTI_\infty}(\boldsymbol{Y}) &\coloneqq \min \boldsymbol{Y} \\
h_{tTMTI_\infty}(\boldsymbol{Y}) &\coloneqq \min(Y_1 + \mu I(\beta^{-1}(1,m)(Y_1) > \alpha), \ldots, Y_m + \mu I(\beta^{-1}(m,1)(Y_m) \geq \alpha)) \\
h_{rtTMTI_\infty}(\boldsymbol{Y}) &\coloneqq \min(Y_1, \ldots, Y_K).
\end{aligned}$$

It is immediate that both h_{TMTI_∞} and h_{rtTMTI_∞} satisfy Condition (C3), as the mapping $\boldsymbol{x} \mapsto \min \boldsymbol{x}$ is weakly increasing in every coordinate. To see that h_{tTMTI_∞} satisfies Condition (C3), note that for fixed $i \in \mathcal{I}$ it holds that $\beta^{-1}(i, m+1-i)$ is strictly increasing, thus making $x \mapsto x+\mu I(\beta^{-1}(i, m+1-i)(x) > \alpha)$ a weakly increasing mapping and therefore also making h_{tTMTI_∞} a weakly increasing mapping.

A.5 Proof of Theorem 4

Assume without loss of generality that $\mathcal{I} = \{1, \dots, m\}$ and $p_1 \leq \cdots \leq p_m$ and denote by $\mathbb{P}_{h(\boldsymbol{Y}^{\mathcal{J}})}$ the CDF of $h(\boldsymbol{Y}^{\mathcal{J}})$. The adjusted p-value for the test of some H_i is the maximal p-value across all intersection hypotheses in the closure of $\{i\}$ in $\mathcal{I}$, i.e.

$$p_i^\star \coloneqq \max_{\mathcal{J} \in \{i\}_{\mathcal{I}}^*} \mathbb{P}_{h(\boldsymbol{Y}^{\mathcal{J}})} \circ h(\boldsymbol{Y}^{\mathcal{J}}).$$

Let $\mathcal{J}_{\{i\}_{\mathcal{I}}^*}^k$ denote the set of all sets in $\{i\}_{\mathcal{I}}^*$ of size k. Then

$$\operatorname*{argmax}_{\mathcal{J} \in \mathcal{J}_{\{i\}_{\mathcal{I}}^*}^k} \mathbb{P}_{h(\boldsymbol{Y}^{\mathcal{J}})} \circ h(\boldsymbol{Y}^{\mathcal{J}}) = \begin{cases} \{m-k, \dots, m\}, & \text{if } i \geq m-k \\ \{i, m-k+1, \dots, m\}, & \text{else} \end{cases}$$

by Lemma 2. Let $\overline{\{i\}_{\mathcal{I}}^*} \coloneqq \{\{i, m\}, \{i, m-1, m\}, \dots, \{i, \dots, m\}, \{i-1, \dots, m\}, \dots, \{1, \dots, m\}\}$. Then, by Lemma 2

$$p_i^\star = \max_{\mathcal{J} \in \overline{\{i\}_{\mathcal{I}}^*}} \mathbb{P}_{h(\boldsymbol{Y}^{\mathcal{J}})} \circ h(\boldsymbol{Y}^{\mathcal{J}}).$$

Thus, the adjusted p-value for any hypothesis H_i can be obtained in $|\overline{\{i\}_{\mathcal{I}}^*}| = m - 1$ steps[2]. However, note for any i that $|\overline{\{1\}_{\mathcal{I}}^*} \cap \overline{\{i\}_{\mathcal{I}}^*}| = i - 1$. Therefore, the number of steps required to obtain an adjusted p-value for all hypotheses is $\sum_{i=1}^m (m-i) = \frac{1}{2} m(m-1)$, as claimed.

A.6 Proof of Lemma 3

Assume without loss of generality that $\mathcal{J}_1 = \{1, \dots, m_1\}$ and $\mathcal{J}_2 = \{1, \dots, m_2\}$. Then $m_1 < m_2$. Let $L \coloneqq \operatorname{argmin}_{j \in \mathcal{J}_1} Y_j^{\mathcal{J}_1}$. Then

$$Z^{\mathcal{J}_1} = \beta(L, m_1 + 1 - L)(P_{(L)}) < \beta(L, m_2 + 1 - L)(P_{(L)}),$$

which implies that $\min \boldsymbol{Y}^{\mathcal{J}_2} < \min \boldsymbol{Y}^{\mathcal{J}_1}$ and therefore $\mathbb{P}(\min \boldsymbol{Y}^{\mathcal{J}_1} < x) < \mathbb{P}(\min \boldsymbol{Y}^{\mathcal{J}_2} < x)$, which is the claimed result.

[2]Disregarding the first step, as a marginal p-value for the test of H_i is already supplied.

B Further simulation studies

B.1 An investigation of the robustness of the TMTI CDFs against different dependency structures

In this section we investigate how well the analytical expression of the TMTI CDFs under the null distribution derived in Section 3.1 under an i.i.d. assumption approximates the actual CDF of the TMTI statistics under different dependency structures.

In the following, we let $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ be an m-dimensional random vector with coordinate means of zero, where $\mathrm{diag}(\Sigma) = (1 \ \ldots \ 1)$. We then calculate p-values $\boldsymbol{P}$ as

$$\forall i \in \{1, \ldots, m\}: \qquad P_i \coloneqq 2(1 - \Phi(|X_i|)),$$

where $x \mapsto \Phi(x)$ is the CDF of a $N(0,1)$ distribution. This ensures that each P_i is uniform on $(0,1)$ and that the dependency structure of $\boldsymbol{P}$ is fully determined by the covariance matrix Σ. We consider three structures of Σ:

1. Equicorrelated tests, where $\Sigma_{i,j} = \rho I(i \neq j) + I(i = j)$, for all i, j and some $\rho \in (0,1)$.
2. Block-diagonal tests, where
$$\Sigma = \begin{pmatrix} \Sigma_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \Sigma_g \end{pmatrix}$$
for some $g < m$ and $\Sigma_1, \ldots, \Sigma_g$ are themselves equicorrelated with parameter ρ.
3. Autoregressive tests, where $\Sigma_{i,j} = \rho^{|i-j|}$.

The first point above can happen in a scenario like the simulation study performed in B.2 where we combine T-tests performed on independent variables but where the standard error is estimated on the basis of a number of covariates. It is unlikely to that the correlation between the tests is high, but it is none-zero. Nevertheless, we try this scenario even for large values of ρ in order to investigate what happens under extreme dependencies.

The second point in the above represents a scenario in which we have performed multiple tests within g groups or individuals that are independent from one another, but where tests performed within the same group are not independent. The last point represents, for example, a design where the tests are spatially correlated, where dependence is highest for neighboring plots.

We perform the experiments with $m = 200$ tests and different values of ρ. For the block-diagonal experiment we set $g = 40$, corresponding to 40 groups of five tests each. In each experiment we bootstrap the CDFs of TMTI$_\infty$, tTMTI∞ with $\tau = 0.05$ and rtTMTI$_\infty$ with $K = 5$ and $K = 1$ both under an assumption of i.i.d. tests and under the actual dependency structure. For comparison, we have also included the Cauchy Combination Test and Harmonic Mean p-value tests. We then plot calibration curves, i.e., the curve $x \mapsto (\text{actual } \mathrm{CDF}(x), \text{i.i.d. } \mathrm{CDF}(x))$. If the i.i.d. CDF is robust against departures from independence then this curve will lie exactly on the diagonal of the unit square. If the i.i.d. CDF is conservative it will lie above the diagonal of the unit square and if it is anti-conservative it will lie below the diagonal of the unit square. The results are presented in Figure 5 and again in Figure 6 where we have zoomed in on the square $(0, 0.1) \times (0, 0.1)$. From the figures, we see that weak dependencies generally do not affect the CDFs of the TMTI statistics by much but stronger dependencies have a large effect on the CDFs. This is similar

to what we see for the Cauchy Combination Test and Harmonic Mean p-value statistics. Although both of these are claimed to be robust against dependence (Liu and Xie, 2020; Wilson, 2019), we see that both of these also become anti-conservative in their lower tails when there is sufficiently strong dependence. The Harmonic Mean p-value appears to generally be more anti-conservative than the TMTI statistics, while the Cauchy Combination Test performs slightly better than rtTMTI$_\infty$ with $K = 5$, but worse than rtTMTI$_\infty$ with $K = 1$. In particular, we see that when dependencies are strong, TMTI$_\infty$ is very anti-conservative in its lower tail implying a loss of Type I error control. However, its truncated variants, tTMTI$_\infty$ and rtTMTI$_\infty$, are less affected by the dependencies and are only slightly anti-conservative in many cases with strong dependencies. In particular, rtTMTI$_\infty$ with $K = 1$, corresponding to only using the smallest p-value, is very robust against most dependencies and, in contrast to the other CDFs presented, never anti-conservative. Only in the equicorrelated experiment when correlations are strong does it deviate from the unit square diagonal and here it is conservative. We also note that the CDFs of all statistics are conservative at their upper tails. However, it is generally only interesting to see how conservative or anti-conservative the i.i.d. approximations are in the regions around typical significance level as conservativeness or anti-conservativeness here can be the difference between making a Type I or II error and making no error.

These experiments suggest that rtTMTI$_\infty$ with $K = 1$ is robust against departures from independence and is thus reasonable to use when making no assumptions on the dependency structure of the p-values in question. In contrast, the other TMTI statistics can be used under some dependency structures if we believe that the dependencies are not too strong, and the truncated TMTI$_\infty$ variants are generally more robust against dependency than TMTI$_\infty$.

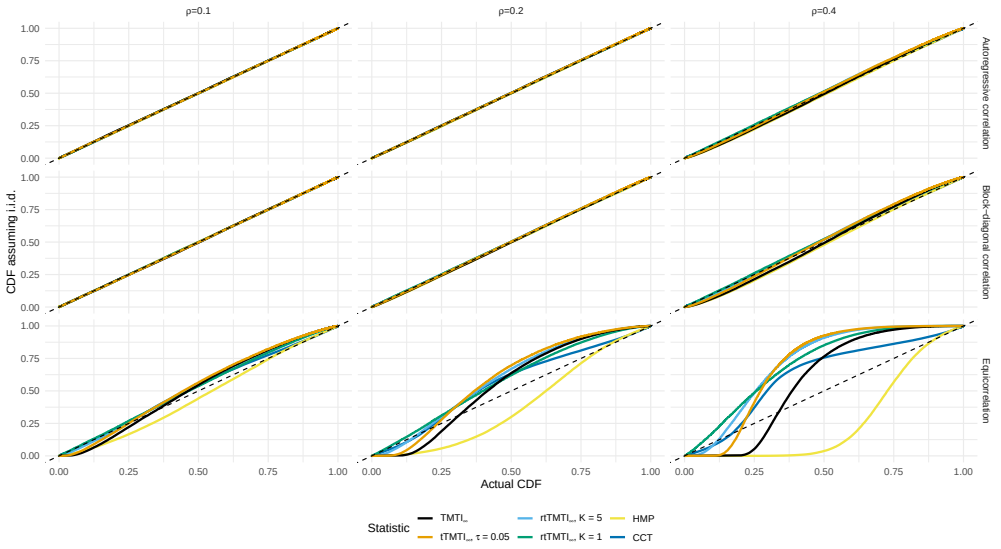


Figure 5: Calibration curves for the CDFs of different TMTI statistics under an i.i.d. assumption versus different dependency structures. CCT = Cauchy Combination Test, HMP = Harmonic Mean p-value.

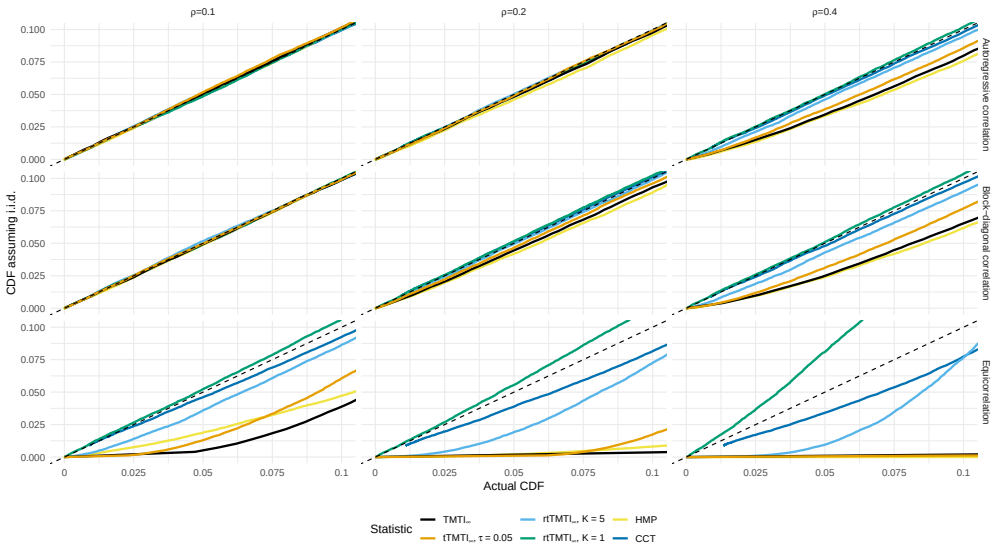


Figure 6: Calibration curves for the CDFs of different TMTI statistics under an i.i.d. assumption versus different dependency structures, zoomed to only show the region $(0, 0.1) \times (0, 0.1)$. CCT = Cauchy Combination Test, HMP = Harmonic Mean p-value.

B.2 An example of applying the TMTI to non-independent data

In this section, we give an example of how the TMTI can be applied to non-independent data in a particular setting. We consider a variation of the simulation study conducted in Section 4, where the marginal p-values now come from dependent T-scores instead of independent Z-scores. Furthermore, we investigate the power of this procedure under different alternatives by means of simulation. Throughout this section, we consider a significance level of $\alpha = 0.05$.

We consider the following setup: let $\boldsymbol{X}$ be an $n \times g$-dimensional binary matrix of full rank satisfying that every row-sum of $\boldsymbol{X}$ equals one and every column-sum of $\boldsymbol{X}$ equals k, for some $k \in \mathbb{N}$. Fix a vector $\mu \in \mathbb{R}^g$ and let $\epsilon \sim N_n(0, I_n)$ be a random variable, where I_n is the $n \times n$ identity matrix . We then define

$$\boldsymbol{W} \coloneqq \boldsymbol{X}\mu + \epsilon.$$

The interpretation of this experiment is that we have recorded k observations of some random variable, W, in m different groups to obtain a total of $N = mk$ samples. By altering the number of non-zero elements in μ we control the number of groups that affect the outcome, $\boldsymbol{W}$. As the power of any test in this scenario is directly associated with the magnitude of the coefficients, μ, we will only consider constant values of μ. That is, the elements in μ which we allow to be non-zero, will all be equal. We then consider the global hypothesis $H_0 : \bigcap_{i=1}^{m}(\mu_i = 0)$, which is the hypothesis that the outcome W is not affected by any of the m groups. The goal is now to estimate the power of the TMTI, i.e., the quantity $\mathbb{P}_{H_A}(\text{reject } H_0)$, under different alternative hypotheses H_A, i.e., different μ vectors. For every alternative hypothesis considered here, we will compute the p-value under four tests from the TMTI family; TMTI_∞ and its truncated and rank truncated versions (with $\tau = 0.05$ and $K = 5$), and TMTI_1. For TMTI_1, we do not consider any truncated variants, as we expect these to be roughly equal (per Figure 1).

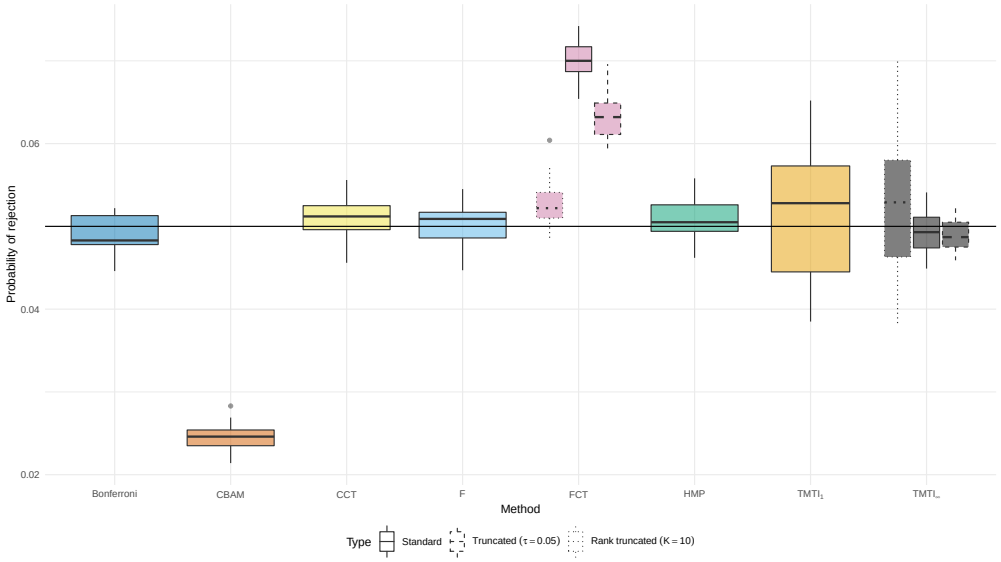


Figure 7: Estimated sizes of the included tests in B.2, i.e., the probability of rejecting the global null hypothesis when it is true. The Truncated Product Methodand Rank Truncated Product Methodare denoted as the truncated FCT and rank truncated FCT, respectively.

For comparison, we consider a number of combination-based test procedures. For this, we will compute the marginal T-tests of the hypotheses $H_i : \mu_i = 0$ for $i = 1, \dots, m$ under the joint model. Then, we will apply the following procedures; the Cauchy Combination Test (Liu and Xie, 2020); the Harmonic Mean p-value (Wilson, 2019); the Compound Bonferroni Arithmetic Mean (Vovk and Wang, 2020); a Bonferroni correction; and the Rank Truncated Product Method. Additionally, we include the F-test, as this would often be the natural choice of test in this particular setup. However, the F-test offers less flexibility compared to a combination test, e.g., when used in a Closed Testing Procedure (see Section 5), as this would require fitting 2^m linear models.

In order to calculate p-values for the TMTI statistics, we estimate the γ functions under H_0 by employing a bootstrapping scheme: For $i = 1, \dots, 10^5$ we first define $\tilde{\boldsymbol{W}} := \boldsymbol{W} - (X^T X)^{-1} X^T \boldsymbol{W} X^T$. That is, $\tilde{\boldsymbol{W}}$ is $\boldsymbol{W}$ with the group means subtracted. This ensures that each group of $\tilde{\boldsymbol{W}}$ has mean zero. We then construct $\boldsymbol{W}_i$ by resampling $\boldsymbol{W}$ uniformly with replacement in order to introduce variation. We then compute the relevant marginal T-test statistics, compute p-values and finally output the TMTI statistic. We then use the empirical CDFs of the bootstrapped TMTI statistics. In Figure 7 we show the simulated sizes of all included tests. From this we conclude that all tests have approximately the correct size or lower, except for the Fisher Combination Test and the Truncated Product Method (both of which assume independence of the p-values), and these are therefore left out from further simulations. One can apply the same bootstrapping scheme as described above to obtain valid p-values for both of these tests. However, we do refrain from doing that here. We note also that the Rank Truncated Product Method has a slightly increased Type I error, but it is sufficiently little that this may be attributed to chance. The increased variance in the estimates of the Type I errors for the TMTI statistics is due to the critical values of these being estimated by bootstrapping.

Figure 8 contains the results of a simulation with $N_{\text{false}} \in \{10^0, \dots, 10^3\}$. Generally, the results are

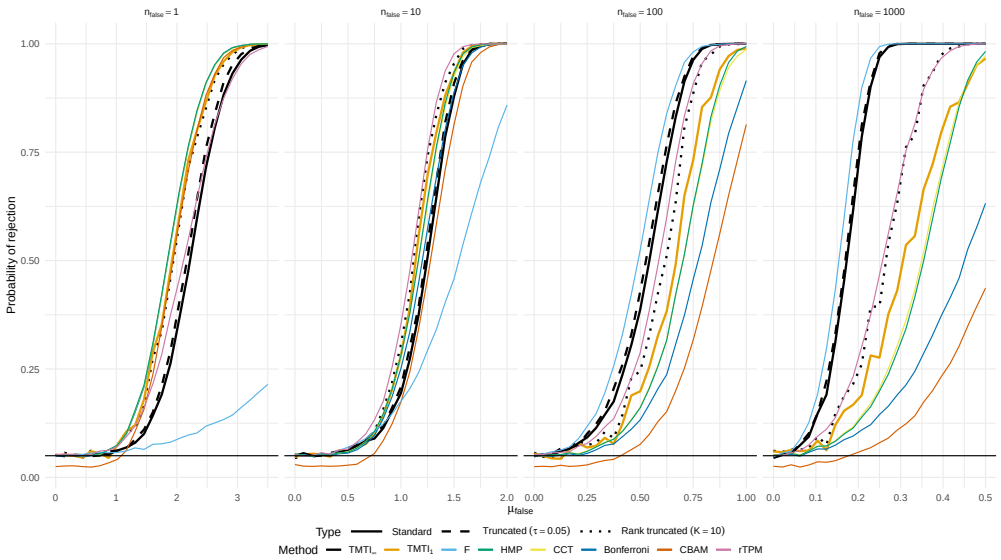


Figure 8: Power curves for different TMTI tests and competing methods. Generally, TMTI_∞ and tTMTI_∞ work well in all settings, while TMTI_1 and rtTMTI_∞ performs best when signals are sparse.

similar to those displayed in Figure 2: TMTI_∞ and tTMTI_∞ both perform well in all settings, although not as well as some methods when $N_{\text{false}} = 1$, but still better than the F-test. When $N_{\text{false}} > 1$, we generally find that TMTI_∞ and tTMTI_∞ perform either as good or better than the best of the competing methods. TMTI_1 and rtTMTI_∞ have similar performance: when there are less than ten false hypotheses, these perform on par or better than with the best of the competing methods, but when there are multiple false hypotheses, they slightly outperform the Cauchy Combination Test and Harmonic Mean p-value, although they are both outperformed by the F-test, TMTI_∞ and tTMTI_∞.

As in Section 4, the results of these simulations indicate that TMTI_∞ and tTMTI_∞ offer an alternative to current tests that is powerful against a wide range of alternative hypotheses. This is useful if one has no *a priori* knowledge of the sparsity and strength of signals, as well as when employing the tests in a Closed Testing Procedure. If one has *a priori* knowledge, that the signals are sparse and strong, one should rather employ a test such as TMTI_1, rtTMTI_∞, Harmonic Mean p-value, Cauchy Combination Test or a Bonferroni test.

Remark 7. *The above example generalizes to a situation in which we have two sets of covariates,* $\boldsymbol{X}$ *and* $\tilde{\boldsymbol{X}}$*, such that* $\boldsymbol{W} \coloneqq \boldsymbol{X}\mu + \tilde{\boldsymbol{X}}\tilde{\mu} + \epsilon$*, where* $\tilde{\boldsymbol{X}}$ *are covariates we would simply like to adjust for when computing our test, but not variables that are of interest. In this setting, we could also compute the* p*-values corresponding to the tests of* $H_i : \tilde{\mu}_i = 0$*, although we would not care whether they are true or not. Thus, we would select a subset of the* p*-values, say* $\mathcal{J}$*, corresponding to those related to* μ*, and compute our test only for those.*

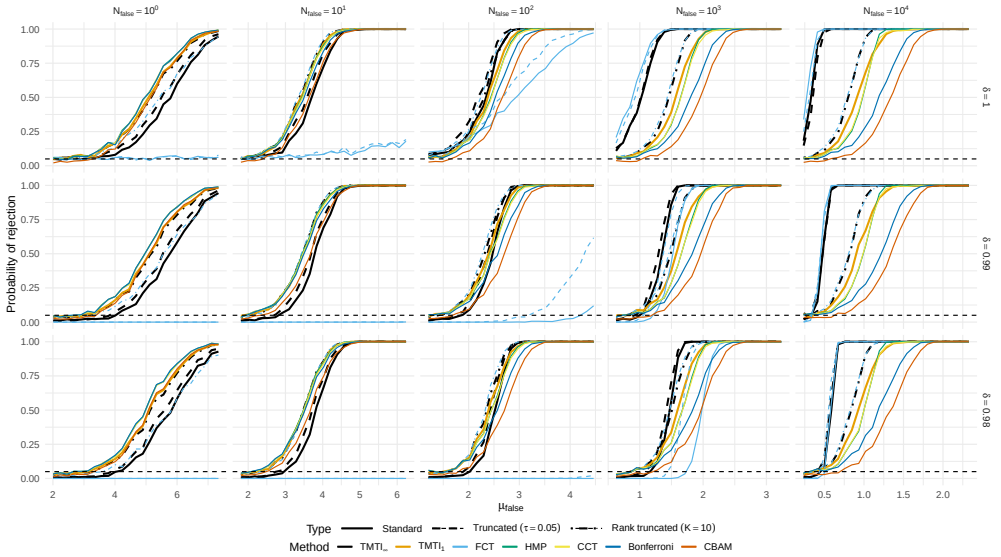


Figure 9: Power curves for different TMTI tests and competing methods, in a scenario where the signal strength is allowed to differ between each false marginal hypothesis. The values of μ_{false} are chosen equidistantly between the two values, which satisfy that a Bonferroni test has either 5% or 99% power to reject the global null hypothesis in a setting with no conservatism.

B.3 The effects of mixed μ values

In this section, we repeat the experiment performed in Section 4, with the change that the μ values are allowed to differ between the false marginal hypotheses. Thus, for N_{false} false hypotheses, we generate p-values by sampling $X_{\text{false},i} \sim N(\mu/i, 1)$ (i.e., the signal with the largest effect has mean μ and the signal with the weakest effect has mean μ/N_{false}), where μ is chosen equidistantly between the values that satisfy that a Bonferroni test has either 5% or 99% power to reject the global null hypothesis in a setting with no conservatism. The results are displayed in Figure 9. The comments to this figure are the same as the comments to Figure 2.

C Table of adjusted p-values for all tests employed in Section 8

State (change)	p-value	Bonferroni	TMTI_∞	rtTMTI_∞/ TMTI_∞	FCT	rTPM/ FCT
GA (-0.323)	0.85628	1.00000	0.87219	0.93682	0.85753	0.93775
AR (-0.777)	0.60282	1.00000	0.87219	0.93682	0.85753	0.93775
AL (-1.568)	0.44008	1.00000	0.85873	0.93682	0.81333	0.93775
NJ (1.565)	0.41998	1.00000	0.85873	0.93682	0.80157	0.93775
NE (1.334)	0.38640	1.00000	0.85873	0.93682	0.78021	0.93775
ND (1.526)	0.36890	1.00000	0.85873	0.93682	0.76813	0.93775
DE (1.374)	0.31162	1.00000	0.85873	0.92675	0.72551	0.92768
MI (2.215)	0.23522	1.00000	0.80175	0.88412	0.66845	0.88500
LA (2.637)	0.20964	1.00000	0.78923	0.88412	0.64602	0.88500
IN (2.149)	0.19388	1.00000	0.78923	0.88412	0.63076	0.88500
WI (2.801)	0.15872	1.00000	0.78923	0.85060	0.59172	0.85144
VA (2.858)	0.14374	1.00000	0.77357	0.84467	0.57388	0.84550
WV (2.331)	0.10026	1.00000	0.68933‡	0.74677	0.51177‡	0.74750
MD (3.339)	0.08226	1.00000	0.68933‡	0.69934	0.48059‡	0.71026
CA (3.777)	0.07912	1.00000	0.68454‡	0.70957	0.47464‡	0.71026
OH (3.466)	0.06590	1.00000	0.62312‡	0.64033	0.44713‡	0.64096
NY (4.893)	0.05802	1.00000	0.58342‡	0.59203	0.42838‡	0.59262
PA (4.303)	0.05572	1.00000	0.58342‡	0.57683	0.42250‡	0.57739
FL (3.784)	0.05490	1.00000	0.58342‡	0.57129	0.42036‡	0.57185
WY (2.226)	0.04678*	1.00000	0.58342‡	0.51259	0.39755‡	0.51308
NM (2.334)	0.04650*	1.00000	0.58342‡	0.51043	0.39671‡	0.51092
CT (3.204)	0.04104*	1.00000	0.55925‡	0.46666	0.37939‡	0.46711
OK (4.181)	0.02036*	0.69224	0.42037‡	0.26549	0.29050‡	0.26573
KY (4.326)	0.00964*	0.32776	0.28899†	0.13524‡	0.21234†	0.13535‡
AZ (4.993)	0.00904*	0.30736	0.27561†	0.12735‡	0.20643†	0.12747‡
ID (2.956)	0.00748*	0.25432	0.23899†	0.10651‡	0.18974†	0.10659‡
TX (5.645)	0.00404*	0.13736	0.17114†	0.05892†	0.14480†	0.05897†
CO (4.326)	0.00282*	0.09588	0.12797†	0.04148*	0.12286†	0.04150*
IA (4.811)	0.00200*	0.06800	0.11058†	0.02958*	0.10453†	0.02961*
NH (4.422)	0.00180*	0.06120	0.10121†	0.02666*	0.09939†	0.02667*
NC (7.265)	0.00002*	0.00068*	0.00346*	0.00346*	0.00843*	0.00064*
HI (5.550)	0.00002*	0.00068*	0.00346*	0.00346*	0.00843*	0.00064*
MN (6.421)	0.00002*	0.00068*	0.00346*	0.00346*	0.00843*	0.00064*
RI (5.094)	0.00001*	0.00034*	0.00198*	0.00198*	0.00551*	0.00044*

Table 2: State-wise changes in mathematics achievements from 1990 to 1992 and the p-values for the corresponding T-tests. Values that are significant at $\alpha = 0.05$ are marked with an asterisk. Hypotheses that can be rejected while controlling the k-FWER at $k = 2$ are marked with a †.

2.2 Intuition

Generally speaking, different tests have different levels of power depending on the alternative hypotheses the null is tested against. Absent a uniformly most powerful test, one therefore chooses a test based on a pre-conceived notion of what the alternative hypothesis is – e.g., how strong the expected effect is. In joint hypothesis testing, the problem of selecting a test is magnified: 'the alternative hypothesis' is not a single alternative hypothesis but a combination of many alternative hypotheses. Thus, selecting a test for a particular joint hypothesis must be done based on how many marginal hypotheses are false but also *how* these are false (i.e., the alternative distribution of each marginal p-value). Some tests have high power against alternatives where there are many (potentially weak) signals, while others work well against alternatives where signals are sparse and strong.

Selecting a joint hypothesis test can be difficult without prior knowledge of the composition of true and false hypotheses. Minimum-p based methods (e.g., Bonferroni) and other methods which are heavily affected by only the smallest p-values (e.g., CCT) work well against sparse alternatives with strong effects but not against dense alternatives with weak effects. In contrast, methods that aggregate the existence of many effects (e.g., FCT) fail when signals are sparse.

With TMTI, we set out to construct a joint hypothesis test capable of adapting to any alternative hypothesis without prior knowledge of it. TMTI accomplishes this (to some degree; see Paper **A**, Section 4) by transforming each p-value separately by different sigmoid functions (i.e., β transformations). Consider for example p-values $P_1, \ldots, P_m$. Each order statistic $P_{(i)}$ is transformed by a function, which has a steep slope in a small neighborhood around the expected value $\mathbb{E}P_{(i)} = i/(m+1)$ under the joint null. If some of the p-values are from false hypotheses, these will be smaller than expected, meaning that the transformed p-values become very small. If only the smallest p-value is from a false hypothesis, we expect that the minimum (whether local or global) is the first of the transformed p-values. However, if more p-values are from false hypotheses, we expect that the minimum lies further along the sequence of transformed p-values. In this sense, we may (informally) think of choosing the minimum of the transformed sequence of p-values, rather than from the p-values themselves, as a way of letting the test statistic adapt to the number of false hypotheses.

2.3 Shortcuts for non-exchangeable distributions

In Paper **A**, Theorem 4 we assumed that the underlying distribution of p-values was exchangeable; this assumption can be further relaxed. We assumed exchangeability to ensure that the CDFs of the TMTI test statistics depend

only on the set $\mathcal{J}$ through its cardinality $|\mathcal{J}|$. In other words, exchangeability ensures that the critical value depends only on $|\mathcal{J}|$ and α. That is, for any test statistic constructed as $Z = h(F_{(1)}(P_{(1)}), \ldots, F_{(m)}(P_{(m)}))$, where $F_{(1)}, \ldots, F_{(m)}$ and h satisfy Conditions (C1) through (C3) of Paper **A**, Lemma 2, it suffices that the critical value of Z depends only on the level, α, and the size of the index set $\mathcal{J}$.

Some test statistics satisfy the property that the critical value of a rejection set depends only on the size of the set and the significance level without assuming exchangeability. For example, the tail of the CCT statistic is robust against departures from independence when the underlying test statistics are jointly Gaussian (Liu and Xie, 2020). Similarly, the methods proposed in Vovk and Wang (2020) are valid under any dependence structure.

2.4 Examples of tests satisfying the conditions of Paper A, Lemma 2

Below, we give examples of test statistics that satisfy the shortcuts of Paper **A**, Lemma 2 when constructed as $Z = h(F_{(1)}(P_{(1)}), \ldots, F_{(m)}(P_{(m)}))$. This list is by no means exhaustive, and some of these are shown to satisfy CTP shortcuts elsewhere. Thus, we go through these simply as an exercise in how one can show that a test statistic satisfies the shortcuts of Paper **A**, Lemma 2.

Proposition 1. *Let $i \in [m]$. The following tests can be constructed as $h(F_{(1)}(P_{(1)}), \ldots, F_{(m)}(P_{(m)}))$ and satisfy the conditions of Paper **A**, Lemma 2.*

1. *The Fisher Combination Test (Fisher, 1992):*

$$F_{(i)}(x) = -2\log(x) \quad \textit{and} \quad h(\boldsymbol{x}) = \sum_{i=1}^{m} x_i.$$

2. *The Kolmogorov generalized f-mean tests (Vovk and Wang, 2020): For $r \neq 0$*

$$F_{(i)}(x) = x^r \quad \textit{and} \quad h(\boldsymbol{x}) = \left(\sum_{i=1}^{m} x_i \right)^{1/r},$$

and for $r = 0$

$$F_{(i)}(x) = x \quad \textit{and} \quad h(\boldsymbol{x}) = \prod_{i=1}^{m} x_i^{1/m}.$$

3. *The Cauchy Combination Test (Liu and Xie, 2020):*

$$F_{(i)}(x) = \tan(\pi(0.5 - x_i)) \quad \textit{and} \quad h(\boldsymbol{x}) = \sum_{i=1}^{m} \omega_i x_i,$$

4. *The Higher Criticism test (Donoho and Jin, 2004):*[1]

$$F_{(i)}(x) = \frac{\sqrt{m}(i/m - x)}{\sqrt{x(1-x)}} \quad \textit{and} \quad h(\boldsymbol{x}) = \max_{1 \le i \le \alpha_0 m} x_i,$$

for any $\alpha_0 \in (0, 1)$.

5. *The (unweighted) Lévy combination test (Wilson, 2021):*

$$F_{(i)}(x) = \Phi^{-1}((1+x)/2) \quad \textit{and} \quad h(\boldsymbol{x}) = m^{-2} \sum_{i=1}^{m} x_i^{-2},$$

where Φ^{-1} *is the quantile function of the* $\mathcal{N}(0, 1)$ *distribution.*

Proof. We prove one by one that the above-listed test statistics satisfy the monotonicity conditions stated in Paper **A**, Lemma 2. For the convenience of the reader, we restate the monotonicity conditions here:

$$\forall j \in [m]\ \forall x \in \mathcal{X}\ \forall \epsilon \ge 0: \qquad F_{(j)}(x) \le F_{(j)}(x+\epsilon) \tag{C1}$$
$$\forall j \in [m-1]\ \forall x \in \mathcal{X}: \qquad F_{(j)}(x) \ge F_{(j+1)}(x) \tag{C2}$$
$$\forall \boldsymbol{x} \in \mathcal{X}^m\ \forall \boldsymbol{\epsilon} \in \mathbb{R}_+^m: \qquad h(\boldsymbol{x}) \le h(\boldsymbol{x}+\boldsymbol{\epsilon}) \tag{C3}$$

$$\forall j \in [m]\ \forall x \in \mathcal{X}\ \forall \epsilon \ge 0: \qquad F_{(j)}(x) \ge F_{(j)}(x+\epsilon) \tag{C1'}$$
$$\forall j \in [m-1]\ \forall x \in \mathcal{X}: \qquad F_{(j)}(x) \le F_{(j+1)}(x) \tag{C2'}$$
$$\forall \boldsymbol{x} \in \mathcal{X}^m\ \forall \boldsymbol{\epsilon} \in \mathbb{R}_+^m: \qquad h(\boldsymbol{x}) \ge h(\boldsymbol{x}+\boldsymbol{\epsilon}) \tag{C3'}$$

Furthermore, we recall that Conditions (C1), (C2) and (C3) or (C1'), (C2') and (C3') together ensures that Paper **A**, Lemma 2 holds if small values are critical for the test statistic. Conversely, Conditions (C1), (C2) and (C3') together, or (C1'), (C2') and (C3) together, ensures that Paper **A**, Lemma 2 holds if large values are critical for the test statistic. If the chosen F functions are identical for all $i \in [m]$, we note that conditions (C2) and (C2') are trivially satisfied, and can therefore be disregarded. In that case, we refer to any $F_{(i)}$ as F.

1. The Fisher Combination Test: F satisfies Conditions (C1') by monotonicity of the logarithm. Furthermore, h trivially satisfies Condition (C3). As large values are critical for FCT, this concludes the proof.

[1] It is generally difficult to compute critical values for this test without resorting to bootstrap methods. It is therefore often suggested that one exchanges $h(\boldsymbol{x})$ with $\tilde{h}(\boldsymbol{x}) = \max_{i: i \le \alpha_0 m,\ x_i > 1/m} x_i$ (see, e.g., Donoho and Jin, 2015). Furthermore, this test is ill-suited to be used by itself in a CTP, as the critical values are only asymptotically valid. However, it can be used as the local test on higher layers of a CTP.

2. The Kolmogorov generalized f-mean tests: For $r \neq 0$, the mappings $x \mapsto x^r$ and $\boldsymbol{x} \mapsto \left(\sum_{i=1}^m x_i\right)^{1/r}$ are strictly increasing on $(0,1)$. Thus, (C1) and (C3) are satisfied when $r \neq 0$. Similarly, $x \mapsto x$ and $\boldsymbol{x} \mapsto \prod_{i=1}^m x_i^{1/m}$ are strictly increasing. Thus, (C1) and (C3) are also satisfied when $r = 0$. As small values are critical for these tests, the proof is complete.

3. The Cauchy Combination Test: The mapping $x \mapsto \tan(\pi(0.5 - x))$ is strictly decreasing on $(0,1)$, as the tangent function is strictly increasing on $(-\pi/2, \pi/2)$. Thus, Condition (C1') is satisfied. Condition (C3) is trivially satisfied. As large values are critical for CCT, the proof is complete.

4. The Higher Criticism of Donoho and Jin: Let $\alpha_0 \in (0,1)$. We note that for any $i \in [m]$

$$\frac{\mathrm{d}}{\mathrm{d}x} F_{(i)}(x) = -\frac{(m-2i)x + i}{2\sqrt{m}(x(1-x))^{3/2}}.$$

The numerator and denominator are strictly positive for $x \in (0,1)$. Thus, Condition (C1') is satisfied. It is furthermore easy to see that Condition (C2') is satisfied. Condition (C3) is satisfied, as the max operator is weakly increasing. As large values are critical for the Higher Criticism test, this completes the proof.

5. The unweighted Lévy combination test As Φ^{-1} is a quantile function it is weakly increasing. Additionally, since $x \mapsto (1+x)/2$ is strictly increasing, Condition (C1) is satisfied. Furthermore, the mapping $x \mapsto x^{-2}$ is strictly decreasing for $x \in (0,1)$ and thus Condition (C3') holds. As large values are critical for the Lévy combination test, the proof is complete. □

2.5 Approximating the CDF of TMTI statistics

In general, it is not trivial to compute p-values for the TMTI tests. For the case when $n = \infty$, we provide in Paper **A** analytic methods for computing these p-values, but these fail when the number of tests is sufficiently large (say, $m \geq 100$). Instead, we primarily use a bootstrapping scheme to compute p-values, but although it is simple, it is computationally expensive. In this section, we argue that it is possible to approximate the p-values for the TMTI tests with $n = \infty$ in a simple manner that does not inflate the Type I error for a pre-specified significance level α.

2.5.1 TMTI and truncated TMTI

First, we argue that the lower α-tail of the CDF of TMTI_∞ and tTMTI_∞ can be approximated by the CDF of a $\beta(1, m')$-distribution for an m' that is

logarithmically polynomial in m. In other words, the lower tail of a TMTI statistic with $n = \infty$ behaves approximately as the minimum of $m' < m$ independent, uniform random variables.

The intuition for this approximation is as follows: it is well known that the minimum of m i.i.d. $U(0,1)$ random variables follows a $\beta(1,m)$-distribution. We also know that transformed p-values $Y_1, \ldots, Y_m$ all follow a $U(0,1)$ distribution under the global null. However, the transformed p-values are not independent. Although we do not know the dependence structure of $(Y_1, \ldots, Y_m)$, we can imagine that dependence has the effect of lumping together some of the Y variables. Thus, we hypothesize that drawing the variables $Y_1, \ldots, Y_m$ is (somewhat) equivalent to drawing only m' independent $U(0,1)$ variables and then repeating some of them, with some level of noise added. If that is the case, the minimum of $Y_1, \ldots, Y_m$ is roughly $\beta(1,m')$-distributed. This intuition applies only to TMTI_∞ and not directly to its truncated versions. However, we show empirically below that the tail of tTMTI_∞ also behaves approximately as a $\beta(1,m')$ variable.

Throughout this section, let us generically denote by $\gamma : [0,1] \to [0,1]$ the CDF of any TMTI statistic (whether truncated or not) when $n = \infty$. When testing at a predetermined level α by using an approximation of γ, say $\hat{\gamma}$, some consideration must go into the choice of $\hat{\gamma}$. A good approximation of γ – i.e., one that rejects and fails to reject the same hypotheses as γ – should satisfy the following:

(i) For all $x \in [0, \gamma^{-1}(\alpha)]$, it holds that $\hat{\gamma}(x) \leq \alpha$.

(ii) For all $x \in (\gamma^{-1}(\alpha), 1]$, it holds $\hat{\gamma}(x) > \alpha$.

These two points are illustrated in Figure 2.1: If $\hat{\gamma}$ only satisfies point (ii), the curve $x \mapsto (\gamma(x), \hat{\gamma}(x))$ will enter the yellow-shaded region. Here, any hypothesis not rejected by γ will not be rejected by $\hat{\gamma}$ either. However, there exist hypotheses that are rejected by γ but not by $\hat{\gamma}$. Thus, $\hat{\gamma}$ gives valid p-values at level α but has lower power than γ. If $\hat{\gamma}$ satisfies points (i) and (ii), the curve moves only through the green-shaded areas, meaning that $\hat{\gamma}$ and γ reject and fail to reject the exact same set of hypotheses. If point (ii) is not satisfied, the curve moves through the red-shaded region, implying that there exist hypotheses that are rejected by $\hat{\gamma}$ but not γ. In this case, $\hat{\gamma}$ does not yield valid p-values at level α.

We can construct an approximation of γ that satisfies point (ii) by setting $\hat{\gamma}(x) \coloneqq \beta(1,m')(x)$ where

$$m' \coloneqq \min\left\{ m'' > 1 \mid \beta(1,m'')(\gamma^{-1}(\alpha)) \geq \alpha \right\}. \tag{2.1}$$

In principle, we should consider integer-valued m'', to be in line with the intuition outlined above, but as the mapping $m \mapsto \beta(1,m)(x)$, for fixed $x \in (0,1)$, is smooth and strictly increasing, we allow m to be a non-integer for

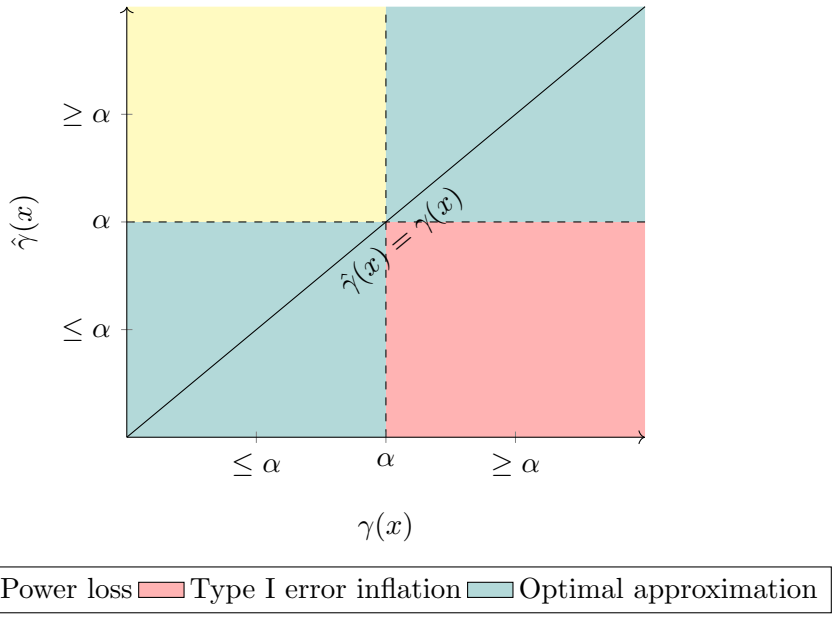


Figure 2.1: Illustration of valid approximations of γ. Any approximation $\hat{\gamma}$ such that the curve $x \mapsto (\gamma(x), \hat{\gamma}(x))$ moves only through the green-shaded areas will reject and fail to reject the exact same hypotheses as γ. Any curve that does not enter the red-shaded area is valid at level α (i.e., controls the Type I error) but has less power to reject a false hypothesis than γ if it moves through the yellow-shaded area.

the added flexibility. The approximation in Equation (2.1) is valid at level α because the regularized incomplete beta function is strictly increasing on $(0, 1)$. Thus, Equation (2.1) implies that $\hat{\gamma}(x) > \alpha$ for all $x \in (\gamma^{-1}(\alpha), 1]$. The solution to Equation (2.1) can be expressed as

$$m' = \frac{\log(1-\alpha)}{\log(1-\gamma^{-1}(\alpha))}.$$

This does not immediately yield an easy-to-use approximation of γ because $\gamma^{-1}(\alpha)$ is computationally expensive to calculate. However, as we shall see in the following simulation studies, the solution to Equation (2.1) is approximately polynomial in the logarithm of m. That is, there exists α-dependent constants $c_{0,\alpha}, c_{1,\alpha}, c_{2,\alpha}$ such that $m' \approx c_{0,\alpha} + c_{1,\alpha} \log m + c_{2,\alpha} (\log m)^2$ for m large enough (say, $m \geq 100$). To estimate these coefficients, we bootstrap $\gamma^{-1}(\alpha)$ 50 times for TMTI_∞ and tTMTI_∞ ($\tau = \alpha$), with $\alpha \in \{0.01, 0.05, 0.10\}$ and $m \in \{2^7, \ldots, 2^{20}\}$, with 10^4 samples per bootstrap. We then compute m' using the bootstrapped estimates of $\gamma^{-1}(\alpha)$. Although the quantiles $\gamma^{-1}(\alpha)$ are not stochastic themselves, there is some stochasticity in m' due to $\gamma^{-1}(\alpha)$ being estimated by bootstrap. The results are displayed in Figure 2.2, where we have fitted quadratic polynomials $m' \approx c_{0,\alpha} + c_{1,\alpha} \log m + c_{2,\alpha} (\log m)^2$,

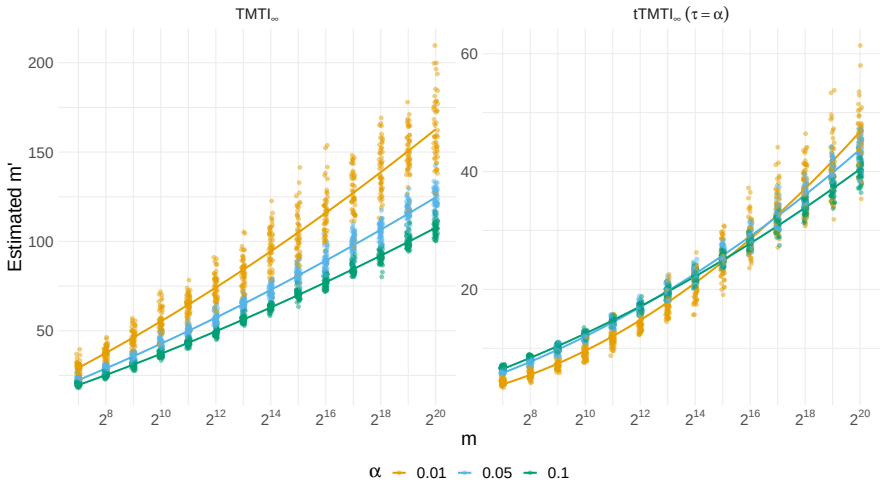


Figure 2.2: Estimates of m' (see Equation (2.1)) based on bootstrap estimates of $\gamma^{-1}(\alpha)$. Each point is based on $B = 10^4$ bootstrap samples. The overlaid lines are quadratic polynomials, $m' \approx c_{0,\alpha} + c_{1,\alpha} \log m + c_{2,\alpha}(\log m)^2$. The estimates of $c_{0,\alpha}, c_{1,\alpha}, c_{2,\alpha}$ are displayed in Tables 2.1a and 2.1b. Left: TMTI_∞. Right: tTMTI_∞ with $\tau = \alpha$. Note the different scales on the y-axis for left and right. Points have been jittered slightly along to x-axis to increase readability.

(a) TMTI_∞

	$c_{0,\alpha}$	$c_{1,\alpha}$	$c_{2,\alpha}$
$\alpha = 0.01$	-20.3067	8.6029	0.3322
$\alpha = 0.05$	-17.3381	7.0808	0.2284
$\alpha = 0.10$	-14.5814	6.0949	0.1965

(b) tTMTI_∞ $(\tau = \alpha)$

	$c_{0,\alpha}$	$c_{1,\alpha}$	$c_{2,\alpha}$
$\alpha = 0.01$	0.4305	-0.7003	0.2917
$\alpha = 0.05$	-2.6523	0.8844	0.1777
$\alpha = 0.10$	-3.2488	1.3951	0.1267

Table 2.1: Estimated coefficients for the $\beta(1, c_{0,\alpha} + c_{1,\alpha} \log m + c_{2,\alpha}(\log m)^2)$ approximations of TMTI_∞ (a) and tTMTI_∞ (b). The estimates are based on the data displayed in Figure 2.2.

which we see fit the data well. The estimated values of $c_{0,\alpha}$, $c_{1,\alpha}$ and $c_{2,\alpha}$ are displayed in Tables 2.1a and 2.1b.

We evaluate the performance of the approximations $\hat{\gamma}$ in two ways. First, we bootstrap γ functions to high precision ($B = 10^6$ bootstrap samples) for $m \in \{10^2, \ldots, 10^6\}$ and consider the curves $(\gamma(x), \hat{\gamma}(x))$ for $x \in [0, \gamma^{-1}(2\alpha)]$; the results are displayed in Figure 2.3. Here, we see that the approximations

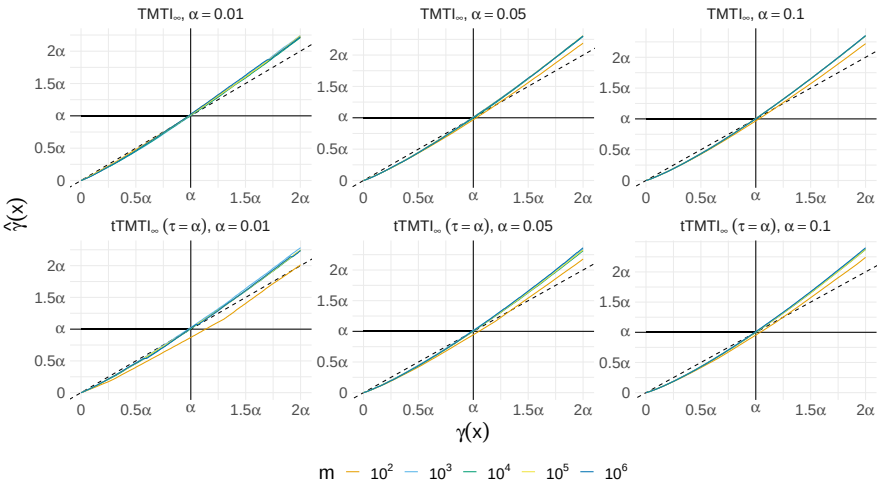


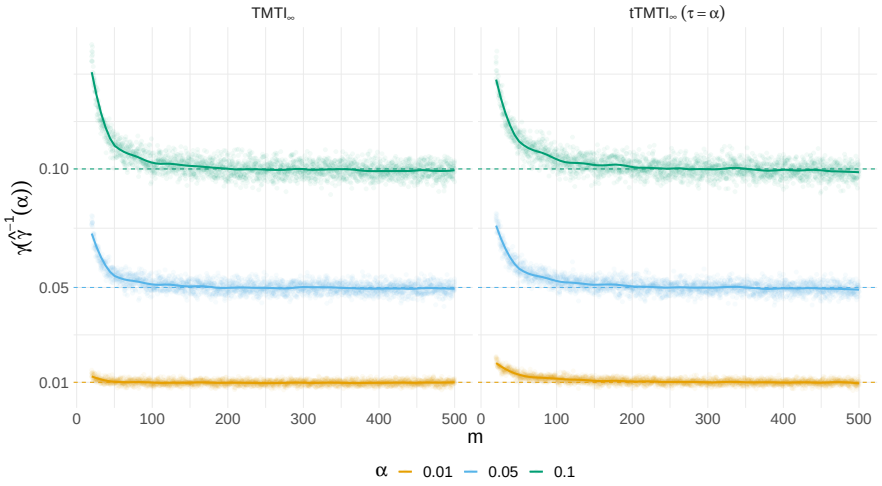
Figure 2.3: Estimated curves $(\gamma(x), \hat{\gamma}(x))$ for $x \in [0, \gamma^{-1}(2\alpha)]$, where $\hat{\gamma}(x) := \beta(1, c_{0,\alpha} + c_{1,\alpha} \log m + c_{2,\alpha}(\log m)^2)(x)$ and $c_{0,\alpha}, c_{1,\alpha}, c_{2,\alpha}$ are, for each configuration, as displayed in Table 2.1. Each estimate of γ is based on $B = 10^6$ bootstrap samples. Generally, every $\hat{\gamma}$ is almost equal to γ for $x \in [0, \gamma^{-1}(\alpha)]$ (except for tTMTI_∞, $\alpha = 0.01$) but becomes larger than γ for $x > \gamma^{-1}(\alpha)$. The dashed line is the identity line $x \mapsto x$.

are almost equal to the respective γ functions for $x \in [0, \gamma^{-1}(\alpha)]$, in all cases except for tTMTI_∞ with $m = 100$ and $\alpha = 0.01 = \tau$. Additionally, we generally see that $\hat{\gamma}(x) \geq \gamma(x)$ for $x > \gamma^{-1}(\alpha)$. This implies that using $\hat{\gamma}$ instead of γ leads to the same rejections and non-rejections as when using γ. Furthermore, we see that p-values obtained by $\hat{\gamma}$ have the same interpretation as those obtained by γ, provided that the p-values are below α.

Second, we estimate the Type I error when rejecting at level α using $\hat{\gamma}$ instead of γ, i.e., the quantity $\gamma(\hat{\gamma}^{-1}(\alpha))$. We do this for small m (here, $m \in \{15, \ldots, 500\}$ with each point repeated five times) and large m (here, $m \in \{2^9, \lfloor 2^{9.5} \rfloor, \ldots, 2^{20}\}$ with each point repeated 50 times) separately. The results are displayed in Figures 2.4a and 2.4b, respectively. These figures show that the approximations generally do not increase the Type I error, except when m is sufficiently low ($m < 100$ for TMTI_∞ and $m < 200$ for tTMTI_∞).

Overall, we find that it is possible to approximate the lower α-tail of the TMTI_∞ and tTMTI_∞ CDFs well by using the CDF of a $\beta(1, c_{0,\alpha} + c_{1,\alpha} \log m + c_{2,\alpha}(\log m)^2)$-distribution, with values of $c_{0,\alpha}, c_{1,\alpha}, c_{2,\alpha}$ listed in Table 2.1. We have tested these approximations for a wide range of m values and found the approximations to work well for all tried $m \geq 100$ or $m \geq 200$ for TMTI_∞ and tTMTI_∞, respectively. Thus, it is reasonable to assume that these approximations also work well for any m between those tested. This yields a considerable speedup, as one no longer needs to bootstrap γ prior to computing the TMTI test.

(a) Small m.



(b) Large m.

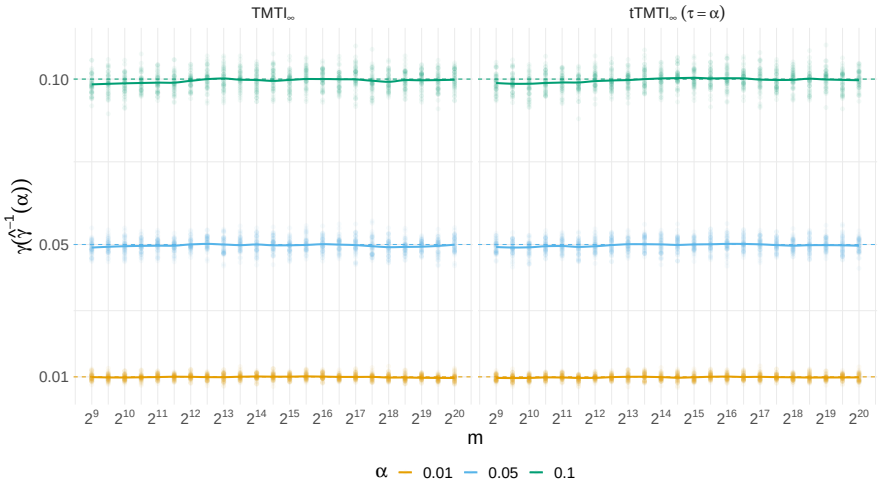


Figure 2.4: Estimates of the Type I error when rejecting at significance level $\hat{\gamma}^{-1}(\alpha)$ instead of α, as described in Section 2.5.1. If $\hat{\gamma}$ is a good approximation of γ in a neighborhood around α, then $\gamma(\hat{\gamma}^{-1}(\alpha)) \approx \alpha$, meaning that the Type I error is still controlled when using $\hat{\gamma}$ instead of γ. This is generally the case when $m \geq 100$ for TMTI_∞ and $m \geq 200$ for tTMTI_∞. (a): $m \in \{15, \ldots, 500\}$. (b): $m \in \{2^9, \lfloor 2^{9.5} \rfloor, \ldots, 2^{20}\}$. In (b), the x-axis is shown on $\log_2$-scale. Overlaid lines are loess smoothers.

2.5.2 Rank truncated TMTI

In this section, we argue that the CDF of rtTMTI_∞ is approximately invariant to the number of independent tests for sufficiently small K. That is, for

$K, m_1, m_2 \in \mathbb{N}$ with $K < m_1 < m_2$, we argue that for $K \leq m_1/4$

$$\forall x \in [0,1]: \quad \gamma_{\infty,K}^{m_1}(x) \approx \gamma_{\infty,K}^{m_2}(x),$$

where $\gamma_{\infty,K}^{m_1}$ and $\gamma_{\infty,K}^{m_2}$ denote the CDFs of rtTMTI for m_1 and m_2, respectively. This implies that one can apply rtTMTI_∞ in a CTP without having to bootstrap the CDF at all layers. Instead, one simply bootstraps the CDF at a single layer, which can then be reused at all other layers, provided that K is sufficiently small relative to the smallest layer it is used at.

Letting $P_{(i)}^{m_1}$ and $P_{(i)}^{m_2}$ denote the i^{th} order statistic of m_1 and m_2 i.i.d. $U(0,1)$ variables, respectively, the intuition behind this approximation is as follows. For $i \leq K$ the transformed p-values $\beta(i, m_1+1-i)(P_{(i)}^{m_1})$ and $\beta(i, m_2+1-i)(P_{(i)}^{m_2})$ have the same marginal distributions. Thus, the distribution of the minima

$$Z^{m_j} \coloneqq \min_{i \leq K} \beta(i, m_j + 1 - i)(P_{(i)}^{m_j}), \quad j \in \{1, 2\},$$

differ only by the dependency structures of

$$(\beta(1, m_1)(P_{(1)}^{m_1}), \ldots, \beta(K, m_1 + 1 - K)(P_{(K)}^{m_1}))$$

and

$$(\beta(1, m_2)(P_{(1)}^{m_2}), \ldots, \beta(K, m_2 + 1 - K)(P_{(K)}^{m_2})).$$

Informally, if K is sufficiently small relative to m_1, then differences in dependency structure have little opportunity to affect the distribution of the minimum.

We verify this empirically by simulating Z^{m_1} and Z^{m_2} for

$$(m_1, m_2) \in \{(10^i, 10^j) \mid i \in \{2, 3, 4\}, j \in \{i+1, \ldots, 5\}\}$$

and

$$K \in \{\lceil 0.05 m_1 \rceil, \lceil 0.10 m_1 \rceil, \lceil 0.25 m_1 \rceil\}.$$

The results are displayed in Figure 2.5. From this figure, we see that no matter the choices of m_1 and m_2, the distributions of Z^{m_1} and Z^{m_2} are approximately equal for $K \leq \lceil 0.1 m_1 \rceil$. For $K = \lceil 0.25 m_1 \rceil$, there appears to be a slight difference between the two distributions.

Next, we consider the curves $(\gamma_{\infty,K}^{m_2}(x), \gamma_{\infty,K}^{m_1}(x))$ for

$$m_1 \in \{10^2, 10^3\}, \quad m_2 = 10^6$$

and

$$K \in \{\lceil 0.05 m_1 \rceil, \lceil 0.10 m_1 \rceil, \lceil 0.25 m_1 \rceil, \lceil 0.5 m_1 \rceil\}.$$

Figure 2.6 contains plots of these curves, zoomed to a typical region of interest, $(0, 0.1)^2$. For both $m_1 = 10^2$ and $m_1 = 10^3$, we see that there is little to no difference between $\gamma_{\infty,K}^{m_1}$ and $\gamma_{\infty,K}^{m_2}$ when $K \leq \lceil 0.10 m_1 \rceil$. When $K = \lceil 0.25 m_1 \rceil$,

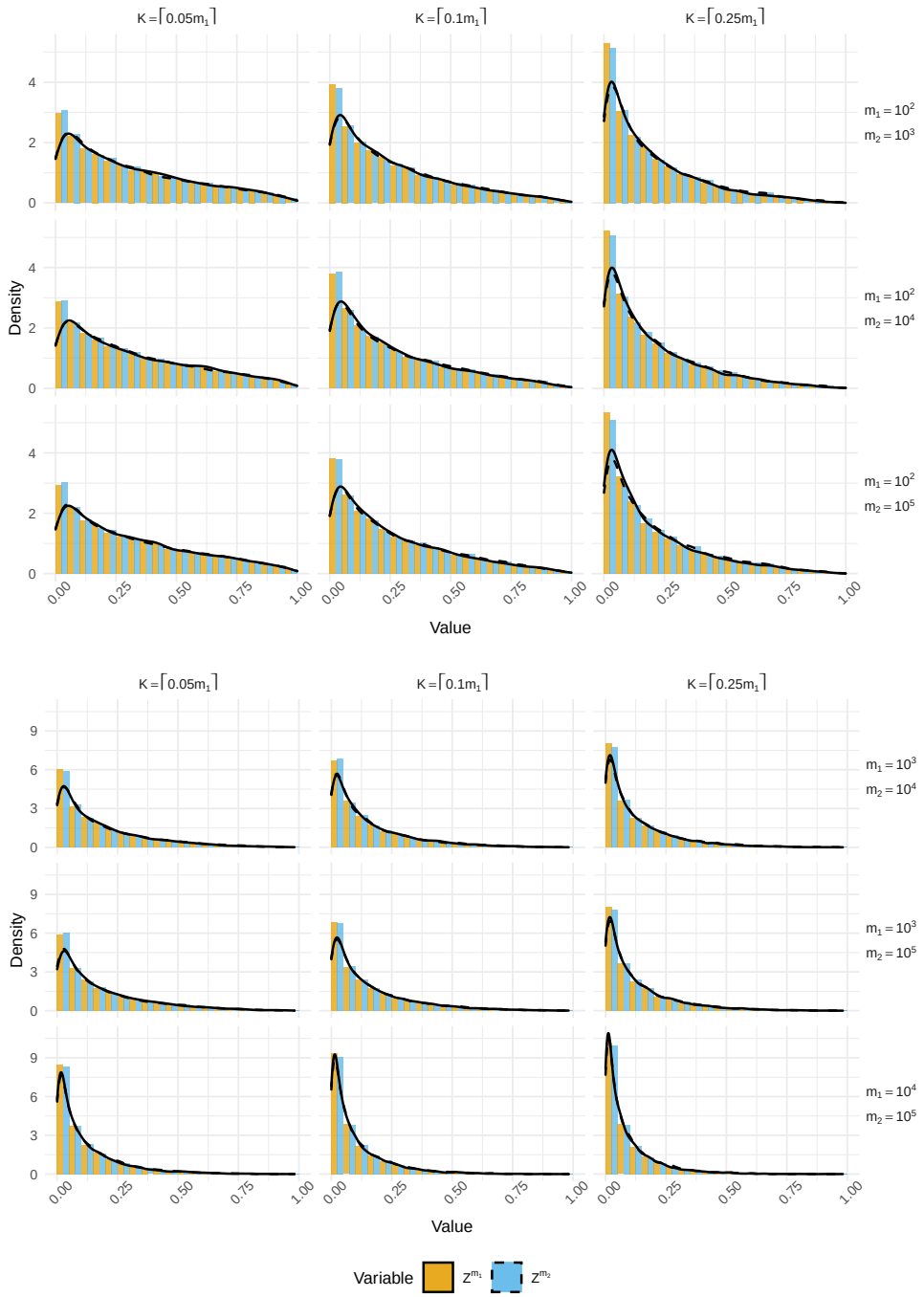


Figure 2.5: Dodged histograms of Z^{m_1} and Z^{m_2} for different choices of $K \in \{\lceil 0.05m_1 \rceil, \lceil 0.10m_1 \rceil, \lceil 0.25m_1 \rceil\}$. The overlaid lines are Gaussian kernel density estimates. Solid line: m_1. Dashed line: m_2. Note that the dashed line is barely visible, indicating that the density estimates are equal for m_1 and m_2. Generally, the distributions of Z^{m_1} and Z^{m_2} are approximately equal for $K \leq \lceil 0.1m_1 \rceil$, with slight differences appearing at $K = \lceil 0.25m_1 \rceil$. Each histogram contains $B = 10^4$ samples.

there is a slight tendency that $\gamma^{m_1}_{\infty,K}(x) > \gamma^{m_2}_{\infty,K}(x)$ for $\gamma^{m_2}_{\infty,K}(x) \leq 0.05$. However, this difference is sufficiently small that it may simply be by chance. For $K = \lceil 0.5m_1 \rceil$, the tendency that $\gamma^{m_1}_{\infty,K} > \gamma^{m_2}_{\infty,K}$ is more pronounced.

In summary, we find that the CDFs $\gamma^{m_1}_{\infty,K}$ and $\gamma^{m_2}_{\infty,K}$ are approximately equal when $K \leq \lceil 0.1m_1 \rceil$ (and possibly $K \leq \lceil 0.25m_1 \rceil$), no matter the choice of $m_1 < m_2$. This implies that using either of the two CDFs will yield the same p-values. For larger K, we find that using $\gamma^{m_1}_{\infty,K}$ will yield slightly larger p-values than using $\gamma^{m_2}_{\infty,K}$. Thus, if applying rtTMTI$_\infty$ with a large K, one can use $\gamma^{m_1}_{\infty,K}$ in place of $\gamma^{m_2}_{\infty,K}$, which will then be conservative.

2.6 On consonance and closed testing

It is commonly accepted that CTPs are better than other multiple testing procedures. However, we have thus far not discussed the question of *what* they are better at. Generally, CTPs have higher power to detect false hypotheses while controlling the FWER (Grechanovsky and Hochberg, 1999) – but that does not mean, that all CTPs have equal power (as seen in Paper **A**, Section 8). For example, consider testing the hypothesis $H_{(1)}$ corresponding to the smallest observed p-value in the closure of m hypotheses. If using, e.g., TMTI$_\infty$ as local test, we can compute the adjusted p-value of $H_{(1)}$ by testing the hypotheses $H_{(1,m)}, H_{(1,m-1,m)}, \ldots, H_{(1,\ldots,m)}$. Now, suppose that $H_{(1)}$ is the only false hypothesis. In that case, it is likely that the TMTI$_\infty$ statistic for the hypothesis $H_{(1,m-i,\ldots,m)}$ will be the first of the transformed p-values. By Paper **A**, Lemma 1, the procedure will then have lower power to detect false hypotheses compared to rtTMTI with $K = 1$ – i.e., a Šidák correction (which has slightly higher power than a Bonferroni correction). In such a case, applying TMTI$_\infty$ in a CTP will likely not reject anything, whereas a Bonferroni correction might. This may seem to conflict with the notion, that CTPs have higher power than non-closed procedures. However, the Bonferroni correction itself is closed: rejecting $H_{(1,\ldots,m)}$ using a Bonferroni test implies that any hypothesis $H_{(1,m-i,\ldots,m)}$ is also rejected by a Bonferroni test.[2]

Not only can we outline examples in which TMTI will not perform as well as other procedures when used in closed testing, we know that there cannot exist *any* scenario in which it performs strongly better than all other CTPs, with respect to FWER control. The reason behind this is that TMTI is *dissonant*:

Definition 2.1. *Let $H^{\mathcal{J}}$ be a joint hypothesis. A CTP is said to be consonant if*

$$H^{\mathcal{J}} \text{ rejected} \implies \exists \mathcal{J}' \subsetneq \mathcal{J} : H^{\mathcal{J}'} \text{ rejected}.$$

A CTP that is not consonant is dissonant.

[2] In fact, using $\boldsymbol{p} \mapsto |\boldsymbol{p}| \min \boldsymbol{p}$ (i.e., the Bonferroni test) as the local test in a CTP is equivalent to Holm-correcting the marginal p-values.

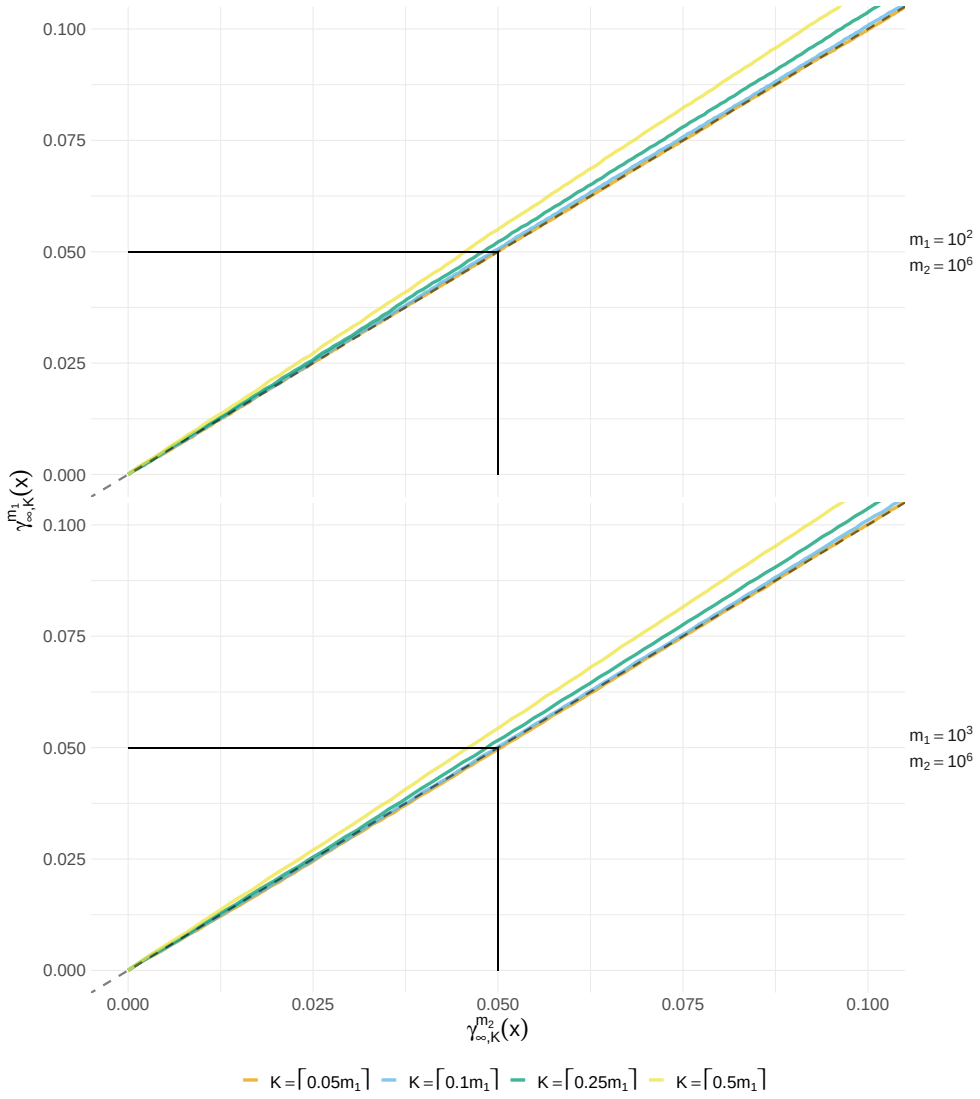


Figure 2.6: Estimated curves $(\gamma^{m_2}_{\infty,K}(x), \gamma^{m_1}_{\infty,K}(x))$, zoomed to the region $(0, 0.1)^2$, for $m_1 \in \{10^2, 10^3\}$, $m_2 = 10^6$ and $K \in \{\lceil 0.05m_1 \rceil, \lceil 0.10m_1 \rceil, \lceil 0.25m_1 \rceil, \lceil 0.5m_1 \rceil\}$. Each curve is based on $B = 10^6$ bootstrap samples. When $K \leq \lceil 0.1m_1 \rceil$, there is no discernible difference between $\gamma^{m_1}_{\infty,K}$ and $\gamma^{m_2}_{\infty,K}$, and when $K = \lceil 0.25m_1 \rceil$ there is a slight difference. When $K = \lceil 0.5m_1 \rceil$, we see that using $\gamma^{m_1}_{\infty,K}$ in place of $\gamma^{m_2}_{\infty,K}$ will be conservative.

Romano et al. (2011) show that any dissonant CTP can be replaced by a consonant CTP which has at least as high power to reject marginal hypotheses as the dissonant CTP. Thus, in any scenario, a direct application of TMTI in a CTP cannot be more powerful than all other CTPs. The reader may then question why we develop such a test. First, while it can not be *more* powerful than the best consonant CTP, it is possible that it is as powerful as the best consonant CTP. Second, as described in Goeman and Solari (2011) and showcased in Paper **A**, Section 8, dissonant procedures provide benefits in other areas than the rejection of marginal hypotheses. For example, we find in Sections 2.7.1 and 2.7.2 that TMTI_∞ performs well with respect to generating confidence sets for the number of false hypotheses in various rejection sets. Thus, the choice of procedure should depend on what one wishes to accomplish. For instance, if the goal is to reject as many false marginal hypotheses as possible with FWER control, and false hypotheses are sparse, then TMTI_∞ is not a good choice, even if signals are strong. Instead, one can employ methods that are invariant to the proportion of false hypotheses – e.g., a Bonferroni correction or, to some extent, CCT. However, if the number of false hypotheses is high relative to the total number of hypotheses, a mixed CTP may be able to identify more false hypotheses (see Paper **A**, Section 8) than the Bonferroni correction. If the goal is not to identify as many false hypotheses as possible, but rather to identify sets in which the relative occurrence of false hypotheses is high, TMTI_∞ is well suited.

2.7 Applications of TMTI to real data

Here, we consider two specific applications of TMTI and closed testing to real datasets. We analyze these datasets, not to draw inference about the subject matter of the data, but to compare the findings when applying different local tests in a CTP. Thus, reported significance levels are invalid if selecting the procedure which yields the best results. The results of each procedure should therefore be interpreted as what we would have concluded, had we only analyzed the data using that procedure.

First, we consider a data set studied in Rasmussen et al. (2022). This data has kindly been supplied by Jacob Agerbo Rasmussen, PhD student at the Section for Hologenomics, University of Copenhagen. Second, we consider an unpublished dataset, supplied by Jaelle Brealey (postdoctoral researcher, Department of Natural History, Norwegian University of Science and Technology) and Morten Tønsberg Limborg (Associate Professor, Section for Hologenomics, University of Copenhagen). This data set was generated as part of the EU Horizon 2020 project HoloFish. In the first data set, from Rasmussen et al. (2022), the number of hypotheses ($m = 569$) is sufficiently low that we can quickly bootstrap the CDFs required to analyze the data using TMTI. For the second data set, we use the approximations proposed in Section 2.5.

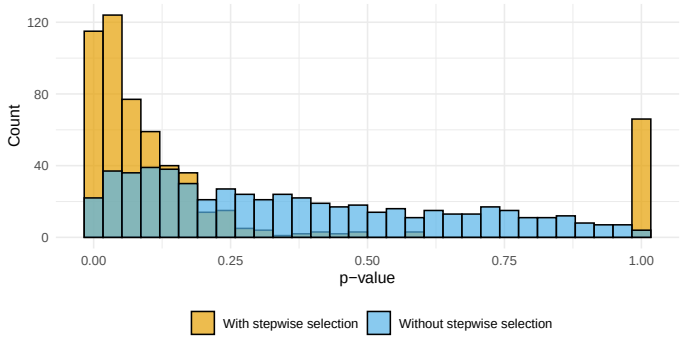


Figure 2.7: Histograms of the p-values from the F-tests described in Section 2.7.1 with (yellow) and without (blue) stepwise model selection. If no variables are selected in the stepwise model selection, we set the p-value to one. The stepwise model selection introduces a downward bias of the p-values, meaning that the marginal Type I error is no longer controlled.

2.7.1 Bacterial Associated Metabolites – data from Rasmussen et al. (2022)

Rasmussen et al. (2022) investigate the association between bacterial Amplicon Sequence Variants (ASVs) and metabolites in commercially available rainbow trout. The data set consists of samples from $n = 26$ rainbow trout, assumed to be independently sampled. Each sample consists of observations of $m = 569$ metabolites $(Y_{1,i}, \dots, Y_{m,i})_{i=1}^n$ and six bacterial ASVs $(X_{1,i}, \dots, X_{6,i})_{i=1}^n$. To identify Bacterial Associated Metabolites (BAMs) – metabolites which are associated with at least one of the six bacterial ASVs – the authors consider, for each $j \in [m]$, the linear regression of the metabolite Y_j onto the six ASVs $X_1, \dots, X_6$, i.e.,

$$\forall i \in [n] : Y_{j,i} = \beta_{j,0} + \sum_{k=1}^{6} \beta_{j,k} X_{k,i} + N_{j,i}, \quad \text{where } (N_{j,i})_{i=1}^n \overset{i.i.d.}{\sim} N(0, \sigma^2).$$

The authors then perform model selection by stepwise minimization of the Akaike Information Criterion (AIC) and perform an F-test for the hypothesis that all remaining coefficients are zero. However, we note that this introduces a downward bias of the computed p-values (see Figure 2.7). We therefore deviate from the procedure of Rasmussen et al. (2022) and instead compute an F-test for the hypothesis $H_{j,0} : \beta_{j,1} = \cdots = \beta_{j,6} = 0$, i.e., we omit the model selection step. Rasmussen et al. (2022) find four BAMs after Bonferroni correcting. No BAMs are found after Bonferroni correcting when omitting the model selection step.

It is reasonable to assume that the metabolites $Y_{1,\cdot}, \dots, Y_{m,\cdot}$ are mutually independent (J. Rasmussen, personal communication). Hence, under any joint null $H_0^{\mathcal{J}} := \bigcap_{j \in \mathcal{J}} H_{j,0}$, the F-tests of the marginal hypotheses $(H_{j,0})_{j \in \mathcal{J}}$ are

independent. Thus, we can thus apply TMTI_∞ to this data. To do this, we bootstrap the relevant CDFs (with $B = 10^5$ bootstrap samples) for joint hypotheses of size $m > 50$ and use their closed-form expressions (see Paper **A**, Section 3.1) when $m \leq 50$.

We first test the global null hypothesis $H_0^{[m]}$ and find that this is rejected at level $\alpha = 0.05$ ($p < 10^{-5}$). Thus, we find evidence that at least one metabolite is associated with at least one ASV. Next, we compute adjusted p-values for all marginal hypotheses and find that none of them can be rejected at level $\alpha = 0.05$. We then compute confidence sets (CSs) for the number of false hypotheses and find that: $\{232, \dots, 569\}$ is a 95% CS for the number of false hypotheses among all 569 hypotheses; $\{52, \dots, 100\}$ is a 95% CS for the bottom 100 hypotheses[3]; $\{11, \dots, 25\}$ is a 95% CS for the bottom 25 hypotheses; and $\{3, \dots, 10\}$ is 95% CS for the bottom 10 hypotheses. Thus, we conclude, for example, that more than every other hypothesis among the bottom 100 is false. This information can, for example, be used to guide a new study in which only those 100 SNPs are investigated, possibly with a larger sample size.

For comparison, we go through the same analysis as above for different choices of local tests and different significance levels, $\alpha \in \{0.01, 0.05, 0.10\}$. Here, we compare to tTMTI_∞ (with $\tau = \alpha$), a Fisher Combination Test (FCT) and a Cauchy Combination Test (CCT). In all cases, we find that no marginal hypotheses can be rejected while controlling the FWER at level α. The results are displayed in Table 2.2. From this table, we see a stark contrast in the results depending on the choice of local test. For example, when $\alpha = 0.05$, TMTI_∞ finds at least 232 false hypotheses among all tested, while CCT is unable to find any. Surprisingly, tTMTI_∞ finds far fewer false hypotheses than TMTI_∞ in any rejection set, despite having power similar to TMTI_∞ in the simulation study in Paper **A**, Section 4. However, in Paper **A**, Section 4 we considered only cases where up to 10 percent of the hypotheses are false – here, more than 40 percent of the hypotheses are false, judging by TMTI_∞. Furthermore, when computing CSs for the number of false hypotheses among all hypotheses, we gradually remove more and more of the smallest p-values as we move along in the procedure. A possible explanation is that tTMTI_∞ does not include p-values that are close to significant and may therefore lose power faster than TMTI_∞ when removing the smallest p-values. Based on this observation, it is natural to ask what the results would have been had we chosen a larger τ. We have therefore repeated the tTMTI_∞ analyses for $\tau \in \{2\alpha, 5\alpha\}$, and the results are displayed in Table 2.3. Interestingly, the performance generally increases when using a larger τ, except in the case where we consider only the bottom ten hypotheses. However, in all cases tTMTI_∞ continues to be outperformed by TMTI_∞. Returning to the results in Table 2.2, we see that FCT also finds far fewer false hypotheses compared to

[3]'Bottom 100' here meaning the hypotheses which gave rise to the smallest 100 p-values.

(a) $\alpha = 0.01$

Method	CS for number of false hypotheses among:			
	All p-values	Bottom 100	Bottom 25	Bottom 10
TMTI_∞	$\{203, \dots, 569\}$	$\{45, \dots, 100\}$	$\{8, \dots, 25\}$	$\{2, \dots, 10\}$
tTMTI_∞	$\{2, \dots, 569\}$	$\{2, \dots, 100\}$	$\{2, \dots, 25\}$	$\{0, \dots, 10\}$
FCT	$\{74, \dots, 569\}$	$\{19, \dots, 100\}$	$\{0, \dots, 25\}$	$\{0, \dots, 10\}$
CCT	$\{0, \dots, 569\}$	$\{0, \dots, 100\}$	$\{0, \dots, 25\}$	$\{0, \dots, 10\}$

(b) $\alpha = 0.05$

Method	CS for number of false hypotheses among:			
	All p-values	Bottom 100	Bottom 25	Bottom 10
TMTI_∞	$\{232, \dots, 569\}$	$\{52, \dots, 100\}$	$\{11, \dots, 25\}$	$\{3, \dots, 10\}$
tTMTI_∞	$\{17, \dots, 569\}$	$\{17, \dots, 100\}$	$\{6, \dots, 25\}$	$\{2, \dots, 10\}$
FCT	$\{84, \dots, 569\}$	$\{24, \dots, 100\}$	$\{0, \dots, 25\}$	$\{0, \dots, 10\}$
CCT	$\{0, \dots, 569\}$	$\{0, \dots, 100\}$	$\{0, \dots, 25\}$	$\{0, \dots, 10\}$

(c) $\alpha = 0.10$

Method	CS for number of false hypotheses among:			
	All p-values	Bottom 100	Bottom 25	Bottom 10
TMTI_∞	$\{247, \dots, 569\}$	$\{56, \dots, 100\}$	$\{12, \dots, 25\}$	$\{4, \dots, 10\}$
tTMTI_∞	$\{42, \dots, 569\}$	$\{35, \dots, 100\}$	$\{8, \dots, 25\}$	$\{3, \dots, 10\}$
FCT	$\{89, \dots, 569\}$	$\{26, \dots, 100\}$	$\{0, \dots, 25\}$	$\{0, \dots, 10\}$
CCT	$\{2, \dots, 569\}$	$\{2, \dots, 100\}$	$\{0, \dots, 25\}$	$\{0, \dots, 10\}$

Table 2.2: Summary of the analyses of the data described in Section 2.7.1 as performed using different local tests at different significance levels; (a): $\alpha = 0.01$; (b): $\alpha = 0.05$; (c): $\alpha = 0.10$. For tTMTI_∞, we set $\tau = \alpha$. 'Bottom 100' refers to the hypotheses that generated the smallest 100 p-values, and similarly for 25 and 10. In all cases, no marginal hypothesis could be rejected while controlling the FWER. No matter the significance level, we see that TMTI_∞ finds the largest number of false hypotheses among any rejection set.

TMTI_∞, but generally more than when using tTMTI_∞. Lastly, using CCT, we do not find any false hypotheses, except when $\alpha = 0.10$, in which case we find at least two false hypotheses among the bottom 100. This is because CCT does not have the power to reject the global hypothesis ($p = 0.08$), except when $\alpha = 0.10$. This is in accordance with Paper **A**, Section 4, where we saw that CCT has low power in scenarios where signals are dense but weak.

To further compare the results, we can also apply methods that aim at estimating the proportion of true hypotheses π_0, and use these to estimate the number of false hypotheses as $(1 - \hat{\pi}_0)m$. Many methods for estimating π_0 have been suggested (see, e.g., Benjamini and Hochberg, 2000; Storey and Tibshirani, 2003; Langaas et al., 2005; Meinshausen and Rice, 2006; Nettle-

Method	CS for number of false hypotheses among:			
	All p-values	Bottom 100	Bottom 25	Bottom 10
$\alpha = 0.01, \tau = 2\alpha$	$\{5, \ldots, 569\}$	$\{5, \ldots, 100\}$	$\{4, \ldots, 25\}$	$\{0, \ldots, 10\}$
$\alpha = 0.01, \tau = 5\alpha$	$\{13, \ldots, 569\}$	$\{13, \ldots, 100\}$	$\{4, \ldots, 25\}$	$\{0, \ldots, 10\}$
$\alpha = 0.05, \tau = 2\alpha$	$\{39, \ldots, 569\}$	$\{33, \ldots, 100\}$	$\{6, \ldots, 25\}$	$\{2, \ldots, 10\}$
$\alpha = 0.05, \tau = 5\alpha$	$\{99, \ldots, 569\}$	$\{39, \ldots, 100\}$	$\{8, \ldots, 25\}$	$\{1, \ldots, 10\}$
$\alpha = 0.10, \tau = 2\alpha$	$\{95, \ldots, 569\}$	$\{41, \ldots, 100\}$	$\{9, \ldots, 25\}$	$\{2, \ldots, 10\}$
$\alpha = 0.10, \tau = 5\alpha$	$\{173, \ldots, 569\}$	$\{48, \ldots, 100\}$	$\{10, \ldots, 25\}$	$\{2, \ldots, 10\}$

Table 2.3: Summary of the analysis of the data described in Section 2.7.1 when using tTMTI_∞ as local test with either $\tau = 2\alpha$ or $\tau = 5\alpha$, for significance levels $\alpha\{0.01, 0.05, 0.10\}$. 'Bottom 100' refers to the hypotheses that generated the smallest 100 p-values, and similarly for 25 and 10.

ton et al., 2006; Pounds and Cheng, 2006; Jiang and Doerge, 2008; Wang et al., 2011). While many of these estimators are constructed to be used in plugin-estimators of the FDR, they can, nevertheless, be applied directly. The methods can only be directly used to estimate the number of false hypotheses among the full set of hypotheses and not among subsets of hypotheses (as one can do with CTPs).[4] Thus, we mainly apply these methods to serve as a sanity check for the results displayed in Table 2.2. As noted in Goeman and Solari (2011), the methods for estimating π_0 that we apply here are best compared to the $1 - \alpha$ confidence sets generated above when $\alpha = 0.5$. These confidence sets are listed in Table 2.4a.

The `R` package `cp4p` (Gianetto et al., 2019) implements a variety of methods to estimate π_0, which we apply to the 569 observed p-values. For methods that require tuning parameters, we use the default settings of the `cp4p` package. The results are displayed in Table 2.4b. Generally, most methods for estimating π_0 agree that a large proportion of the hypotheses are false. Five methods estimate between 50% and 60% of the hypotheses being false, and one method estimates 25% percent of the hypotheses being false. Two methods estimate a very low proportion of false hypotheses (1% and 6%, respectively). The authors in Benjamini and Hochberg (2000) argue that the estimator is conservatively biased, i.e., $\hat{\pi}_0 > \pi_0$, meaning that the corresponding estimator of the number of false hypotheses is biased to be lower than the actual number of false hypotheses. This potentially explains the large difference. The estimator proposed in Wang et al. (2011) depends on three tuning parameters, for which we have used the default settings. However, the default parameters may not be appropriate for this particular data. Indeed, by modifying the tuning parameters slightly, we were able to obtain estimates of π_0 ranging from 0.3 to 0.97, with most estimates falling in the range 0.8 to 0.95 (not

[4]The reason for this is that confidence sets generated by CTPs hold simultaneously for all choices of rejection sets. In contrast, applying the same estimator $\hat{\pi}_0$ to multiple subsets and selecting the one which gives the best results induces a selection bias.

(a)

Method	50% confidence set
TMTI_∞	$\{299, \ldots, 569\}$
tTMTI_∞	$\{202, \ldots, 569\}$
FCT	$\{110, \ldots, 569\}$
CCT	$\{26, \ldots, 569\}$

(b)

Method	$\hat{\pi}_0$	$1 - \hat{\pi}_0$	$(1 - \hat{\pi}_0)m$
Storey and Tibshirani (2003)	0.39	0.61	347.91
Storey et al. (2004)	0.43	0.57	322.77
Langaas et al. (2005)	0.45	0.55	315.18
Jiang and Doerge (2008)	0.46	0.54	307.40
Nettleton et al. (2006)	0.50	0.50	285.50
Pounds and Cheng (2006)	0.74	0.26	150.05
Wang et al. (2011)	0.94	0.06	34.99
Benjamini and Hochberg (2000)	0.99	0.01	7.99

Table 2.4: (a): 50% confidence sets for the number of false hypotheses among all 569 tested hypotheses. For tTMTI_∞ we have used $\tau = 0.05$. (b): The estimated proportion of true null hypotheses $\hat{\pi}_0$ and the corresponding estimates of the proportion and number of false hypotheses, respectively, when applying different methods for estimating π_0. The estimates were generated using the `R` package `cp4p` (Gianetto et al., 2019). Most methods find that somewhere between 40% and 50% of the hypotheses are false, one finds 26% to be false, and two methods estimate considerably lower proportions.

shown). From this, we conclude two things. First, the performance of each estimator varies considerably, just as the performance of each method in Tables 2.2 and 2.4a varied considerably by the choice of local test. Second, most estimators find more false hypotheses than the best performing method in Table 2.4a, TMTI_∞. However, many estimators in Table 2.4b do not provide true confidence bounds, contrary to the confidence sets generated by CTPs.

2.7.2 Single Nucleotide Polymorphism associated gut bacteria – data from HoloFish

In this section, we analyze an unublished data set from the EU Horizon 2020 project HoloFish. As part of this study, nine data sets were generated, each investigating possible associations between Single Nucleotide Polymorphisms (SNPs) and the abundance or presence of nine different bacteria found in the gut microbiome of 463 harvest-aged farmed Atlantic Salmon. Here, we consider just one of these data sets: presence/absence of the bacteria *Photo-*

bacterium iliopiscarium. This data set was selected by Jaelle Braelly (post-doctoral researcher, Department of Natural History, Norwegian University of Science and Technology) as the 'biologically most interesting data set'. The data set consists of 998,475 p-values – one for each SNP tested.

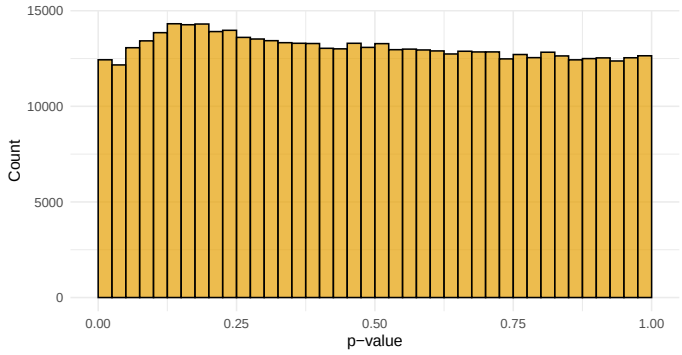
These p-values are unlikely to be independent due to linkage disequilibrium (LD) (see, e.g., Slatkin, 2008), and thus it is unreasonable to apply a TMTI test directly. Since LD is a local phenomenon (Dudbridge and Koeleman, 2003), it is possible to filter the p-values such that the remaining are approximately independent. This filtering is done by only keeping p-values for SNPs that are sufficiently many base pairs apart. Here, it is reasonable to assume independence of p-values corresponding to SNPs that are at least 1,000 base-pairs apart within each chromosome (Shyam Gopalakrishnan, personal communication).[5] Thus, we keep only p-values for SNPs that lie at least 1,000 base-pairs apart, starting with the first observation in each chromosome. This approach has the advantage that we can apply methods, such as TMTI, that rely on independence. However, we may also remove potential findings if the p-values we remove are from false hypotheses. On the other hand, filtering might also increase power if – by chance – the p-values removed are from true hypotheses. In this case, it may be easier to detect the false hypotheses among the reduced set of hypotheses. An alternative approach is to use the full set of p-values and apply a method like CCT, which satisfies the same CTP shortcut as TMTI and does not require independence. However, CCT cannot be applied to this particular data, as there are p-values that are exactly one, and the CCT test statistic is therefore undefined.

After filtering, the data contains 523,196 p-values. An overview of these p-values is provided in Figure 2.8. In non-human GWAS studies, such as this, it is common that signals are sparse and weak, and the histogram in Figure 2.8a is in line with this. Here, there is a slight over-representation of p-values in the left-most half of the histogram, but this appears to consist mostly of near-significant (at level $\alpha = 0.05$) p-values. In the Manhattan plot in Figure 2.8b, we see that there are many signal spikes, but only three of these are strong enough to be rejected by a Bonferroni correction.

As discussed in Section 2.6, CTPs with local tests lacking power in sparse scenarios are unlikely to produce any significant findings with FWER control. Indeed, when applying a Bonferroni (or Šidák) correction, three of the $m = 523{,}196$ adjusted p-values are below the significance level $\alpha = 0.05$. In contrast, no adjusted p-values are below α when using a CTP with either TMTI_∞, tTMTI_∞ ($\tau = 0.05$) or FCT. When applying rtTMTI_∞ ($K = 5$), we can reject two marginal hypotheses. Thus, when the number of false hypotheses is low relative to the total number, the value of such tests is found in their dissonant rejections.

[5] In case this filtering is insufficient to ensure independence, we compare only to other methods that assume independence, so that no method has an unfair advantage over another.

(a) Histogram of the filtered p-values.



(b) Manhattan plot of the filtered p-values.

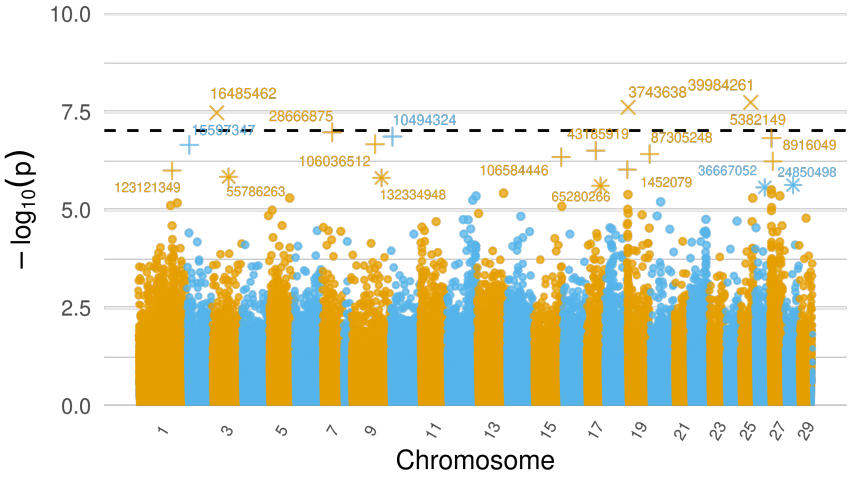


Figure 2.8: An overview of the p-values analyzed in Section 2.7.2. (a): We see that there is an overweight of small p-values, but most of these are larger than the typical 5% significance level. (b): The dotted line represents the Bonferroni significance level on $-\log_{10}$ scale. Points marked by '$\times$' can be rejected with FWER control when Bonferroni (or Šidák) correcting. Points marked by either '$\times$' or '$+$' can be rejected by TMTI_∞ and tTMTI_∞ ($\tau = 0.05$) while controlling the k-FWER at $k = 5$. Points marked by either '$\times$', '$+$' or '$*$' can be rejected by rtTMTI_∞ ($K = 5$) with k-FWER control at $k = 5$. The labels represent the SNP positions relative to their chromosomes.

Method	No. of rejections at k-FWER control:				
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
TMTI_∞	0	3	8	12	14
tTMTI_∞	0	3	10	12	14
rtTMTI_∞	2	9	13	16	19
FCT	0	0	0	0	0

Table 2.5: Attempted methods and the number of hypotheses in Section 2.7.2 that can be rejected by each, when controlling the k-FWER at level $\alpha = 0.05$ for different k. For tTMTI_∞, we used $\tau = 0.05$. For rtTMTI_∞, we used $K = 5$. For comparison, a Bonferroni correction rejects three hypotheses.

First, we find the largest sets that can be rejected with k-FWER for different k. Due to the computational complexity of finding these sets, we do this only up to $k = 5$. The findings are displayed in Table 2.5. From this table, we see that truncated methods fare better than non-truncated methods for controlling the k-FWER for small k.

Changing tack, we compute confidence sets for the number of false hypotheses for pre-specified rejection sets instead. The results are displayed in Table 2.6. Here, we see that TMTI_∞ is generally best at generating narrow confidence sets for the number of false hypotheses in rejection sets. That is, TMTI_∞ finds more false hypotheses than the other methods, when considering at least the 25 smallest p-values. The truncated procedures, tTMTI_∞ and rtTMTI_∞, find very few false hypotheses among the full set of hypotheses. However, when considering smaller subsets, the relative drop in findings is lower for tTMTI_∞ and rtTMTI_∞ compared to TMTI_∞ and FCT.

In summary, Tables 2.5 and 2.6 indicate that TMTI_∞ is generally well suited for finding as many as possible false hypotheses among rejection sets. However, TMTI_∞ is outperformed by rtTMTI_∞ when the considered sets of hypotheses are sufficiently small. Though none of the applied methods can reject as many marginal hypotheses with FWER control as a Bonferroni/Šidák correction, they are useful for finding large rejection sets with many false hypotheses.

2.8 An overview of the R package TMTI

This section contains a discussion of some of the practical considerations that went into the implementation of the TMTI family of tests. In addition, we give a brief description of the R package TMTI (Mogensen, 2021) (available

Method	CS for number of false hypotheses among:			
	All p-values	Bottom 1,000	Bottom 100	Bottom 25
TMTI_∞	$\{17{,}762, \dots, m\}$	$\{156, \dots, 10^3\}$	$\{45, \dots, 10^2\}$	$\{16, \dots, 25\}$
tTMTI_∞	$\{145, \dots, m\}$	$\{145, \dots, 10^3\}$	$\{46, \dots, 10^2\}$	$\{17, \dots, 25\}$
rtTMTI_∞	$\{15, \dots, m\}$	$\{15, \dots, 10^3\}$	$\{15, \dots, 10^2\}$	$\{15, \dots, 25\}$
FCT	$\{947, \dots, m\}$	$\{0, \dots, 10^3\}$	$\{0, \dots, 10^2\}$	$\{0, \dots, 25\}$

Table 2.6: Attempted methods and 95% confidence sets for the number of false hypotheses in different subsets. For tTMTI_∞, we used $\tau = 0.05$. For rtTMTI_∞, we used $K = 5$. Here, $m = 523{,}196$. 'Bottom 1,000' refers to the hypotheses that generated the smallest 1,000 p-values, and similarly for 100 and 25.

on `CRAN` and https://github.com/PhillipMogensen/TMTI).[6,7] This package was developed to accompany Paper **A**. The core of the package consists of nine functions: `TMTI`, `adjust_*`, `TestSet_*`, `TopDown_*` and `kFWER_*`, where `*` means either `TMTI` or `LocalTest`. The `*_TMTI` functions are simply wrappers around the `*_LocalTest` functions, where the local test of choice is a TMTI test. In the `*_LocalTest` functions, the user can supply their own local test. The function `TMTI` computes a single TMTI test given input p-values. The functionalities of the remaining functions are summarized in Table 2.7.

Throughout this section, we compare the computation times of different implementations. All computations were performed using an Apple M1 Pro 3.2GHz CPU. A table summarizing all the comparisons can be found in Table 2.8.

2.8.1 Optimizing the computation of TMTI statistics

In general, computing the TMTI statistic is not problematic when computing a single test. For instance, given a vector of p-values `p`, the test statistic for TMTI (truncated or not) can for $n < m$ be implemented in `R` as:

```
Z_R_loop = function (
    p,  # A pre-sorted, pre-truncated vector of p-values
    n,  # type of minimum to consider
    m   # total number of p-values prior to truncation
) {
    PreviousY = 1
    m_p       = length(p)
```

[6] At the time of submission of this thesis, an updated version of `TMTI` has been submitted to CRAN, but not yet published. See the GitHub repository of the package for the most recent version.

[7] A frozen copy of the GitHub repository `TMTI` is available at https://github.com/PhillipMogensen/TMTI_FrozenCopy. This frozen instance reflects the status of the `TMTI` package at the time this thesis was submitted.

Function	Purpose
`adjust_*`	Computes adjusted p-values for all marginal hypotheses, or optionally only for the marginal hypotheses that can be rejected with FWER control. If `direction = 'binary'` is specified, the function only identifies the number of hypotheses that can be rejected with FWER control, and not the actual p-values.
`TestSet_*`	Tests a user-specified joint or marginal hypothesis in a CTP, i.e., it computes an adjusted p-value for a user-specified set.
`TopDown_*`	Uses a binary search to compute a confidence set for the number of false hypotheses in a user-specified rejection set. This function is named `TopDown_*` purely for historical reasons. When we first implemented it, we did not use a binary search, but instead started from the largest set and iteratively checked smaller and smaller sets. Hence, the name `TopDown_*`
`kFWER_*`	Computes the largest set of marginal hypotheses that can be rejected with k-FWER control for a user-supplied k.

Table 2.7: Purposes of the main functions of the `R` package `TMTI`. All functions are efficiently implemented in `C++`.

```
8       out        = -1
9
10      # This keeps track of the leading n Y_i
11      LeadingY = pbeta(p[2:(n + 1)],
12                       2:(n + 1),
13                       m + 1 - 2:(n + 1))
14
15      for (i in 1:(m_p - n)) {
16          Y = pbeta(p[i], i, m + 1 - i)
17          LeadingMin = min(LeadingY)
18          # If Y is smaller than the previous Y and
19          # smaller than the following n, return Y
20          if ((Y < PreviousY) &
21              (Y <  LeadingMin)) {
22              out = Y
23              break
24          }
25          # If not, update the leading n Y_i
```

```
        PreviousY = Y
        LeadingY = c(LeadingY[-1], pbeta(p[i + n + 1],
                                         i + n + 1,
                                         m + 1 - (i + n + 1)))
    }

    # If no minimum is found, select the minimum
    # of the last n Y_i
    if (out == -1)
        out = LeadingMin

    return(out)
}
```

In the above, `p` is pre-sorted and pre-truncated (if applicable) and `m` is the number of p-values prior to truncation (`m` = `m_p` if no truncation is performed). The general case is handled in the package `TMTI`. If n is sufficiently large, `Z_R_loop` is slower than using a version where Y is computed up-front, because the function `stats::pbeta` is a vectorized `C++` function, and therefore faster than iteratively computing each element in `R`:

```
Z_R_vectorized = function (p, n, m) {
    Y         = pbeta(p, 1:m, m:1)
    PreviousY = 1
    m_p       = length(p)
    out       = -1

    for (i in 1:(m_p - n)) {
        if ((Y[i] < PreviousY) &
            (Y[i] < min(Y[(i + 1):(i + n)]))) {
                out = Y[i]
                break
            }
            PreviousY = Y[i]
    }

    if (out == -1)
        out = min(Y[(m_p - n + 1):m_p])

    return(out)
}
```

Note that if $n = m - 1$, these functions return $\min Y$, and therefore finds the (rt/t)TMTI$_\infty$ statistic. In the case $n = \infty$ (or equivalently $n \geq m - 1$),

it easier and more efficient (because we avoid any explicit R loops and only compute a single minimum) to implement this as:

```
Z_R_infty = function (p, m) {
    m_p = length(p)
    Z   = min(stats::pbeta(p, 1:m, m:1))
    return(Z)
}
```

These implementations are decently fast (see Table 2.8a). For example, when applied to $m = 10^6$ random and pre-sorted p-values, it takes on average 59×10^{-6} seconds to compute the TMTI statistic when $n = 1$ using `Z_R_loop`. When $n = 10^4$, the computation time increases to 0.04 seconds, and at $n = 10^5$ it takes just under half a second. Going up to $n = \infty$, it takes on average 0.15 seconds to compute the TMTI statistic using `Z_R_infty`. While 0.15 seconds is not slow, per se, it is slow when viewed in the context of closed testing, where the computation must be performed possibly billions of times (although some of these computations will be for small m and therefore faster). However, for the case of $n = \infty$, there is little optimization to be done, because `Z_R_infty` relies only on the functions `min` and `stats::pbeta`, both of which are already heavily optimized. However, we can slightly improve the runtime if we implement it in `C++` instead, using the `Rcpp` framework (Eddelbuettel and François, 2011; Eddelbuettel, 2013; Eddelbuettel and Balamuta, 2018). This reduction in runtime is possible because the minimum can be computed in an online fashion, meaning that we can avoid a term that has linear time complexity:[8]

```
double Z_C_infty(NumericVector p, int m) {
    int m_p           = p.size();
    double currentMin = 1;
    double currentY   = 0;
    for(int i = 0; i < m_p; i++) {
        currentY = R::pbeta(p[i],
                            i + 1,
                            m + 1 - (i + 1),
                            true,
                            false);
        if(currentY < currentMin){
            currentMin = currentY;
        }
    }
    return currentMin;
}
```

[8]The reader may also note that this implementation is more memory efficient than `Z_R_infty`.

The above implementation is on average around 7% faster than `Z_R_infty` for $m = 10^6$ (see Table 2.8a). For smaller m, the reduction in computation speed is smaller, but still positive (not shown). In the case of $n < m - 1$, there is a considerable gain in computation speed by rewriting `Z_R_loop` in `C++`, as shown below.[9]

```
1  double Z_C(NumericVector p, int n, int m) {
2      int m_p          = p.size();
3      double Z         = -1;
4      double PreviousY = 1;
5      double Y;
6      std::vector<double> LeadingY;
7      double LeadingMin;
8
9      // Initialize the minimum of the leading n Y_i
10     for (int i = 1; i <= n; i++) {
11         LeadingY.push_back(R::pbeta(p[i],
12                                     i + 1,
13                                     m - i,
14                                     true,
15                                     false));
16     }
17
18     // Iteratively compute Y_i and check if
19     // it is smaller than the leading n Y_i
20     for (int i = 0; i < m_p - n; i++) {
21         Y = R::pbeta(p[i], i + 1, m - i, true, false);
22         LeadingMin = *std::min_element(LeadingY.begin(),
23                                        LeadingY.end());
24         if ((Y < LeadingMin) & (Y < PreviousY)) {
25             Z = Y;
26             break;
27         }
28         LeadingY.erase(LeadingY.begin());
29         LeadingY.push_back(R::pbeta(p[i + 1 + n],
30                                     (i + 1 + n) + 1,
31                                     m - (i + 1 + n),
32                                     true,
33                                     false));
```

[9] For large n and m, the computation time can potentially be slightly reduced by replacing the `vector LeadingY` with a `deque` (double-ended queue) type and line 28 with `LeadingY.pop_front();`, as front-deletion has constant complexity for `deque` types but linear complexity for `vector` types. However, `deque` types have slightly worse performance on some other operations, potentially negating the difference.

```
34          PreviousY = Y;
35      }
36
37      // If no minimum is found, select the
38      // minimum of the last n Y_i
39      if (Z == -1) {
40          Z = LeadingMin;
41      }
42
43      return Z;
44  }
```

The above implementation is almost six times faster than `Z_R_loop` for $m = 10^6$ when $n = 1$, and ten times faster when $n = 10^4$ (see Table 2.8a). In practice, one will rarely use n different from one or infinity, but these implementations allow for it. In addition, TMTI with $n = 1$ does not immediately allow for shortcuts in CTPs (see Paper **A**, Remark 5), although there are situations in which $n = 1$ satisfies a shortcut at some layers (see Section 2.9). Overall, the gain in computational speed from these optimized implementations is therefore small (but not negligible).

2.8.2 Early stopping, search strategies and parallelization

Early stopping: Paper **A**, Lemma 2 implies that for some combination tests, adjusted p-values for m marginal hypotheses can be obtained in quadratically many steps. However, in many cases, practitioners are only interested in presenting adjusted p-values for the marginal hypotheses that can be rejected while controlling the FWER. We can obtain adjusted p-values for only the rejectable hypotheses by employing early stopping. Suppose we are given p-values for m hypotheses and that R of these can be rejected while controlling the FWER using some procedure satisfying the shortcut of Paper **A**. Paper **A**, Lemma 2 implies that adjusted p-values for these R hypotheses can be obtained in at most $m(R + 1)$ steps, and that the R hypotheses that can be rejected will be those corresponding to the R smallest marginal p-values. This is because the adjusted p-value for $H_{(R+1)}$ will be: 1) larger than α, and 2) smaller than the adjusted p-value of any hypothesis $H_{(i)}$, $i > R + 1$. Thus, if starting from the smallest p-value, we can stop the adjustment procedure as soon as we reach an adjusted p-value that is above α, knowing that we have found all hypotheses that can be rejected with FWER control. This can be specified in the `TMTI` package by supplying the argument `EarlyStop = TRUE`.

Search strategies: The method outlined above presumes that we always start by adjusting the smallest p-value first. This does not need to be the case. In particular, there are two other search strategies we can employ. However,

using these strategies, we only identify the number of hypotheses that can be rejected with FWER control.

First, if we reverse the search direction and instead start from the largest p-value, we can stop the procedure the first time we encounter an adjusted p-value that is below α. Once we encounter an adjusted p-value below α, we know that all smaller marginal p-values have adjusted p-values below α as well. This does not alter the computational complexity of the procedure. However, it can be considerably faster than searching from the smallest p-value, if used in combination with early stopping, and if sufficiently many hypotheses can be rejected. This reversed search strategy can be specified in the `TMTI` package by supplying the argument `direction = 'decreasing'`.

Second, we can employ a binary search to identify the smallest unadjusted p-value that is above α. By the same argument as before, this allows us to identify the number of hypotheses that can be rejected with FWER control. In contrast to the `decreasing` strategy, a binary search *does* alter the computational complexity. In particular, employing a binary search allows us to identify the number of hypotheses that can be rejected with FWER control in $\mathcal{O}(m \log m)$ time.[10] A binary search strategy can be specified by supplying the argument `direction = 'binary'`. A simple example of how a binary search can be implemented is given below. The function `TestSet_C`, that is called in line 16, is defined in Section 2.8.3.

```
1  int FWER_set_C (Function LocalTest,
2                  std::vector<double> pvals,
3                  double alpha,
4                  int low,
5                  int high) {
6      int mid;
7      mid = (low + high) / 2;
8
9      // Extract the p-value at index mid
10     std::vector<double> p_mid;
11     p_mid.insert(p_i.end(), pvals[mid]);
12     // Copy pvals and remove index mid
13     std::vector<double> p_copy = pvals;
14     p_copy.erase(b.begin() + mid);
15
16     double p = TestSet_C(LocalTest,
17                          p_mid,
18                          p_copy,
19                          alpha,
```

[10] The adjustment of each p-value has complexity $\mathcal{O}(m)$. The worst-case number of adjusted p-values that need to be calculated in order to identify the smallest adjusted p-value above α is $\log m$ when using a binary search.

```
20                          TRUE);
21
22      if ((low >= high) & (p < alpha)) {
23          return low + 1;
24      } else if ((low >= high) & (p >= alpha)) {
25          return low;
26      } else if (p < alpha) { // Target is left of mid
27          return FWER_set_C(LocalTest,
28                            pvals,
29                            alpha,
30                            mid + 1,
31                            high);
32      } else { // Target is right of mid
33          return FWER_set_C(LocalTest,
34                            pvals,
35                            alpha,
36                            low,
37                            mid);
38      }
39
40      return -1;
41  }
```

Search strategies and confidence sets: Similar to identifying the number of hypotheses that can be rejected with FWER control, we can employ a binary search to improve the computational complexity of computing $1 - \alpha$ confidence sets for the number of false hypotheses in a rejection set S. Searching iteratively for the largest subset of S that cannot be rejected has complexity $\mathcal{O}(m|S|)$ when $|S| < m$. By employing a binary search instead, the complexity is reduced to $\mathcal{O}(m \log |S|)$.[11]

We can also improve the computational complexity in the case when $S = [m]$, if the local test we employ satisfies a mild form of coherence (see, e.g., Romano et al., 2011). In particular, if the non-rejection of $H_{(i,\ldots,m)}$ at level α implies the non-rejection of $H_{(j,\ldots,m)}$, $i < m$, at level α, the computational complexity becomes $\mathcal{O}(\log m)$ instead of $\mathcal{O}(m)$.[12] Note that it is only in the case of $S = [m]$, that the local test needs to satisfy this property in order to gain a computational speedup. For $|S| < m$, the closure principle ensures that the property is satisfied. However, we conjecture that this property is satisfied for most combination tests at reasonable levels of α.

Binary search methods are implemented as default in the functions `TopDown_*`.

[11] S contains $|S|$ hypotheses to test, each of which can be tested in linear time.
[12] S contains m hypotheses to test, each of which can be tested in constant time.

Search strategies and k-FWER control: In Paper **A**, we stated that the largest possible rejection set which controls the k-FWER can be found in cubic time when $k > 1$. However, this result can be improved to $\mathcal{O}(m(\log m)^2)$ time by employing a double binary search. That is, we can employ two binary searches simultaneously: one to look for the largest set S with at least $|S|+1-k$ false hypotheses, and another that determines the number of false hypotheses in each set S (as described above). This is implemented in the functions `kFWER_*` in the `TMTI` package.

Depth-wise parallelization: We generally consider two parallelization schemes: 'depth-wise' and 'breadth-wise'. Suppose we have n_w available workers to parallelize onto. A depth-wise parallelization distributes the work of adjusting the p-value of a single hypothesis – say $H_{(1)}$ – onto n_w workers. That is, to adjust the p-value of $H_{(1)}$, we must test the hypotheses $H_{(1)}, H_{(1,m)}, \ldots, H_{(1,3,\ldots,m)}, H_{(1,\ldots,m)}$. When distributing the task of testing these hypotheses, we generally do so in chunks to allow for early stopping. That is, we split the set

$$\{(1, m), (1, m-1, m), \ldots, (1, \ldots, m)\}$$

into n_c chunks of size $c \geq n_w$,

$$\begin{aligned} c_{1,\alpha} =& \{(1, m), \ldots, (1, m-c, \ldots, m)\}, \\ c_{2,\alpha} =& \{(1, m-c-1, \ldots, m), \ldots (1, m-2c, \ldots, m)\}, \end{aligned}$$

and so forth. We then loop over all chunks, parallelizing the test of the hypotheses $\{H_i\}_{i \in c_j}$ onto n_w workers. Upon collection of the results of one chunk, we evaluate whether any of the hypotheses in that chunk failed to be rejected, and break the computation if that is the case. Parallelizing in chunks adds overhead to the computation because the distribution of work has to happen up to n_c times instead of once, as would be the case without early stopping. Thus, there is a trade-off to consider: using larger chunks means less overhead, but if it is possible to stop the execution early, the amount of wasted computation power increases. A drawback of the depth-wise parallelization scheme is that we generally cannot perform this type of work distribution using `Rcpp`. A framework for parallelization in `Rcpp` – called `RcppParallel` (Allaire et al., 2022) – does exist. However, `RcppParallel` cannot be directly applied for the tasks we consider here, as we always input a user-defined `R` function, `LocalTest`. Thus, we make calls to the `R` API from within `C++`, which is discouraged in threaded code, per the 'Writing R Extensions' manual.[13]

Breadth-wise parallelization: An alternative to depth-wise parallelization is to employ a breadth-wise parallelization. Here, we instead distribute

[13] See https://cran.r-project.org/doc/manuals/r-release/R-exts.html

the task of adjusting the p-values for every hypothesis of interest $H_{(i)}$ onto its own worker. This approach has the benefit, that the task of adjusting the p-value of any one hypothesis can be done entirely within C++. However, a drawback of breadth-wise parallelization, is that we cannot employ binary searches with this parallelization strategy. Both depth-wise and breadth-wise parallel schemes are implemented in the package TMTI, using the option parallel.direction. We compare the different parallelization strategies in Section 2.8.3 when we have discussed different implementations of TestSet_*.

2.8.3 Optimizing the TestSet_* function

The function TestSet_*, which tests a single hypothesis in the closure of all hypotheses, is a workhorse of the TMTI package. As TestSet_* is called in both adjust_* and TopDown_*, it is relevant to optimize it, as this is where the majority of the computation time will be spent. A simple R implementation of this function looks as follows:

```
TestSet_LocalTest = function (
  LocalTest,  # a function defining a local test
  p_subset,   # p-values for the set to test
  p_rest,     # the remaining p-values
  alpha,      # significance level
  EarlyStop   # FALSE/TRUE
) {
    m          = length(p_rest)
    CurrentMax = 0

    p_first = LocalTest(p_subset)
    if ((p_first >= alpha) & EarlyStop)
        return (p_first)

    for (i in m:1) {
        p = LocalTest(c(p_subset, p_rest[i:m]))
        if (p > CurrentMax)
            CurrentMax = p
        if (EarlyStop & (CurrentMax > alpha))
            break
    }

    return(CurrentMax)
}
```

The above assumes that the vectors p_subset and p_rest[14] are sorted and that the subset consists of the smallest |subset| p-values. We handle the

[14]The p-values of the subset to be tested and the remaining p-values, respectively

general case in the package `TMTI`. This implementation of `TestSet_R` takes on average around half a second to adjust the p-value of $H_{(1)}$ without early stopping when $m = 10^4$ and around 49 seconds when $m = 10^5$ (see Table 2.8b).[15] Profiling `TestSet_R` reveals that the majority of the time is spent allocating and deallocating memory in the step `p = LocalTest(c(p_subset, p_rest[i:m]))`. For example, when $m = 10^5$, around 40 Gb of memory is allocated and deallocated again during the process.[16] The reason for this is that `R` needs to allocate a new vector `c(p_subset, p_rest[i:m])` at every step and then delete it again. To circumvent this issue, we implement the procedure in `C++`:

```
double TestSet_C (
    Function LocalTest,  // a function defining a local test
    std::vector<double> p_subset, // p-values for the set to
                                  // test
    std::vector<double> p_rest,   // the remaining p-values
    double alpha,                 // significance level
    bool EarlyStop                // FALSE/TRUE
) {
    int n  = p_rest.size();
    int n2 = p_subset.size();
    double p;
    double CurrentMax = 0;

    // test H_subset and break if not rejected
    double p_first = *REAL(LocalTest(p_subset));
    if ((p_first >= alpha) & (EarlyStop)) {
        return p_first;
    }

    for (int i = 0; i < n; i++) {
      // Insert the largest remaining p-value immediately
      // after the largest of the p-values in p_subset
      auto it = p_subset.begin() + n2;
      p_subset.insert(it, p_rest.back());
      // Delete the largest p-value in p_rest
      p_rest.pop_back();
      // Compute the test of the current subset
      p = *REAL(LocalTest(p_subset));
      // Update the current max p-value
```

[15] Here, we have used a Bonferroni correction as local test. We use the Bonferroni correction for these comparisons to minimize the time spent on computing the local tests.

[16] At layer m' a vector of m' doubles, which each take up 8 byte, has to be allocated and deallocated. Thus, the total allocation is 8 bytes $\times \sum_{i=1}^{10^5} i \approx 40$Gb.

```
30        if (p > CurrentMax) {
31          CurrentMax = p;
32        }
33        // End the process if we failed to reject and early
34        // stopping is specified
35        if ((p > alpha) & (EarlyStop)) {
36          break;
37        }
38      }
39
40    return CurrentMax;
41  }
```

This implementation of `TestSet_C` is twice as fast as `TestSet_R` when $m = 10^5$ and roughly 50% faster when $m = 10^4$ (see Table 2.8b). The main difference between `TestSet_R` and `TestSet_C`, is that we convert the `R` vectors `p_subset` and `p_rest` to `C++` vectors. These vectors in `C++` can be appended and shortened without copying the entire vector at each iteration.

As noted in Section 2.8.2, a drawback of this `C++` implementation is that it is generally more difficult to parallelize than its `R` counterpart, in which the loop can be directly parallelized. However, it is, easy to call `TestSet_C` in parallel processes. That is, given a sorted vector `my_pvalues`, we can adjust the p-values of the marginal hypotheses $H_{(1)}$ through $H_{(i)}$ in a breadth-wise parallel scheme, e.g., by:

```
p_adjusted = parallel::mclapply (
    1:i,
    function (j) {
        # Compute the test of H_j
        TestSet_C (
            LocalTest = <a function defining a test>,
            p_subset  = my_pvalues[j],
            p_rest    = my_pvalues[-seq(j)],
            alpha     = <significance level>,
            EarlyStop = <FALSE/TRUE>

        )
    },
    # Set the number of workers
    mc.cores = min(i, parallel::detectCores())
)
```

However, the drawback of the above breadth-wise parallelization is that the overall computation is not broken if a worker fails to reject the marginal hypothesis. For instance, if $H_{(1)}$ cannot be rejected, the above will still test $H_{(2)}$

through $H_{(i)}$, even though these are not necessary to compute.[17] However, parallelization can still be performed in chunks, which allows the process to end, when a chunk has been computed in which a hypothesis could not be rejected.

In Table 2.8c, we have benchmarked `TestSet_C` against a depth-wise parallel implementation of `TestSet_R` and a breadth-wise parallel implementation of `TestSet_C`. For comparison, we also include a binary search based on `TestSet_C` for the number of hypotheses that can be rejected with FWER control. Here, we simulated $m \in \{10^3, 10^4\}$ p-values, constructed such that $H_{(1)}$ through $H_{(10)}$ can be rejected by a Bonferroni correction while $H_{(11)}$ cannot.[18] When $m = 10^4$, we see from Table 2.8c that `TestSet_C` takes on average 3.3 seconds to adjust the p-values of $H_{(1)}$ through $H_{(10)}$ when single-threaded. In contrast, it takes on average 2.4 seconds to adjust the same p-values when using a depth-wise parallel implementation of `TestSet_R` (six cores, chunksize $10^4/2$). When using a breadth-wise parallel implementation of `TestSet_C` (six cores, chunksize six), the computation time reduces to just 0.8 seconds. The binary search implementation is in this case faster even than the breadth-wise parallelization of `TestSet_C`, taking just 0.4 seconds on average to complete. However, the binary search only tells us how many hypotheses can be rejected, whereas the other methods also compute adjusted p-values.

The implementation of `TestSet_LocalTest` in `TMTI` is based on `TestSet_R` when depth-wise parallelization is specified. In all other cases, the implementation is based on `TestSet_C`.

2.8.4 Bookkeeping

There is some overlap in the tests performed when carrying out a CTP using the implementations we discussed above. Thus, there is some wasted computational power. For sparse data – i.e., data where few hypotheses can be rejected – this waste is negligible. Similarly, for both sparse and dense data, the waste is negligible if we are only interested in identifying the number of hypotheses that can be rejected with FWER control. Thus, there is only a non-negligible waste if many hypotheses can be rejected and the goal is to obtain adjusted p-values for these hypotheses.

However, it is not difficult to construct a function that adjusts all p-values while maintaining the bookkeeping. The bookkeeping should preferably be done entirely within `C++`, as the overhead of maintaining the bookkeeping in `R` is likely so high that performance deteriorates with bookkeeping compared to without bookkeeping. Below is a simple example of how bookkeeping can be handled entirely within `C++`:

[17] In practice, however, these will take up very little computation time, as they will likely fail to be rejected early in the procedure and therefore broken early internally.

[18] Here, we have again used a Bonferroni correction as the local test.

```
1  std::vector<double> FullCTP_C (
2      Function LocalTest,   // a function defining a local test
3      Function f,           // a call to TestSet_C
4      std::deque<double> p  // a deque of p-values
5  ) {
6      std::vector<double> BottomTrees;
7      std::vector<double> TopTree;
8      std::vector<double> out;
9      double max;
10     int m = p.size();
11
12     for (int i = 0; i < m - 1; i++) {
13         // Test all hypotheses that lie in an overlap:
14         TopTree.push_back(*REAL(LocalTest(p)));
15
16         double p_current = p.front();
17         p.pop_front();
18         // Get the largest p-value for H_{(i + 1)} among
19         // hypotheses that do not lie in an overlap:
20         BottomTrees.push_back(*REAL(f(p_current, p)));
21     }
22     TopTree.push_back(p[0]);
23
24     // Combine the overlapping hypotheses with non-overlapping:
25     for (int i = 0; i < m; i++) {
26         max = *std::max_element(TopTree.begin(),
27                                 TopTree.begin() + i + 1);
28         if (BottomTrees[i] > max) {
29             out.push_back(BottomTrees[i]);
30         } else {
31             out.push_back(max);
32         }
33     }
34
35     return out;
36 }
```

In the code above, the function `f` will be a call to `TestSet_C` (as defined in Section 2.8.3) with `p_subset` = `p_current` and `p_rest` = `p`. In the `TMTI` package, `FullCTP_C` has been wrapped in a convenience function `CTP_LocalTest` and slightly modified to allow for early stopping. Visually, lines 12-22 of `FullCTP_C` correspond to computing all hypotheses connected by a south-east facing arrow in Paper **A**, Figure 3 and all hypotheses connected by a south-west facing arrow separately. In lines 25-33 of `FullCTP_C`, the south-west facing branches

of the test tree are combined with the south-east facing branch to obtain the adjusted p-values.

In Table 2.8d, we have compared the median time required to adjust $m = 10^2$ and $m = 10^3$ p-values, when using either `FullCTP_C` versus iteratively calling `TestSet_C` (without early stopping), both single-threaded and in parallel. For both $m = 10^2$ and $m = 10^3$, we see that `FullCTP_C` is roughly twice as fast as iteratively calling `TestSet_C`. When $m = 10^2$, `FullCTP_C` is even faster than calling `TestSet_C` in parallel, but this difference disappears when $m = 10^3$. Thus, `FullCTP_C` is useful for smaller m, if one has no intention of parallelizing. For larger m, however, employing a parallel version of `TestSet_C` is still preferable.

2.9 Future outlook

We list here some topics for further research related to what we have presented in this chapter.

Analytical TMTI tests for dependent p-values: In many cases, it is of interest to apply combination tests to dependent data. While some methods exist for this (e.g., CCT), it may also be interesting to derive analytical CDFs of the TMTI statistics in the case of dependent p-values. An approach to solve this, is to consider dependent Z-scores $(Z_1, \ldots, Z_m) \sim N(0, \Sigma)$ with corresponding p-values $P_i \coloneqq 2 \times (1 - \Phi(|Z_i|))$. However, to calculate the CDF of, e.g., TMTI_∞, we need to derive the joint distribution of the order statistics of the absolute Z-scores, i.e., the distribution of $(|Z|_{(1)}, \ldots, |Z|_{(m)})$. To the best of our knowledge, this distribution is not known. Under the additional assumption that the inverse covariance matrix Σ^{-1} is an M-matrix (every off-diagonal element is weakly negative)[19], it is easy to show that the (non-ordered) distribution of $(|Z_1|, \ldots, |Z_m|)$ is Multivariate Totally Positive of order two (MTP_2).[20] Using this fact, along with already existing bounds on order statistics of MTP_2 variables (see, e.g., Sarkar, 1998; Sarkar and Smith, 1986), it may be possible to derive conservative bounds on the TMTI CDFs.

Approximate CTP shortcuts for $n = 1$ TMTI tests: We showed in Paper **A**, Remark 5, that the quadratic shortcuts do not generally apply to TMTI statistics when $n = 1$. However, there are cases in which we can skip some tests. To illustrate this, consider the testing of, e.g., $H_{(i)}$ at the layer where we test all joint hypotheses of size m' that contain $H_{(i)}$. Suppose the

[19] This is, for example, the case for most forms of positive dependence (Colangelo et al., 2005).

[20] If Σ^{-1} is an M-matrix, Z is MTP_2 (Fallat et al., 2017, Section 4.1); MTP_2 is preserved under strictly increasing coordinate-wise transformations (Fallat et al., 2017, Proposition 3.1).

minimum of the transformed p-values is at the first position in the test of $H_{(i,m-m'+2,\dots,m)}$, i.e., that

$$Y_1 \coloneqq \beta(1, m')(P_{(i)}) < \beta(2, m' - 1)(P_{(m-m'+2)}) =: Y_2.$$

Then, the largest p-value among all tests of size m' is obtained when considering the i^{th} p-value combined with the $m' - 1$ largest remaining p-values. Thus, at some layers, it is only necessary to perform a single test. In cases where this shortcut does not apply, it is still possible to use Y_1 as a conservative test statistic. However, it is unclear how conservative this bound is and how often the shortcut applies.

(a) Median time to compute TMTI test statistics.

Function	Setting: $m = 10^6$			
	$n = 1$	$n = 10^4$	$n = 5 \times 10^4$	$n = \infty$
`Z_R_loop`	$0.059 \times 10^{-3}s$	$0.039s$	$0.497s$	–
`Z_R_vectorized`	$150.9 \times 10^{-3}s$	$0.178s$	$0.251s$	–
`Z_C`	$0.011 \times 10^{-3}s$	$0.004s$	$0.184s$	–
`Z_R_infty`	–	–	–	$0.152s$
`Z_C_infty`	–	–	–	$0.141s$

(b) Median time to adjust a single p-value

Function	Setting	
	$m = 10^4$	$m = 10^5$
`TestSet_R`	$0.477s$	$48.89s$
`TestSet_C`	$0.314s$	$25.65s$

(c) Median time to adjust 10 p-values

Function	Setting	
	$m = 10^3$	$m = 10^4$
Single-threaded `TestSet_C`	$0.100s$	$3.315s$
Depth-wise parallel `TestSet_R`	$0.193s$	$2.394s$
Breadth-wise parallel `TestSet_C`	$0.059s$	$0.803s$
Binary search `TestSet_C`	$0.023s$	$0.354s$

(d) Median time to adjust all p-values

Function	Setting	
	$m = 10^2$	$m = 10^3$
Single-threaded `TestSet_C`	$0.067s$	$8.786s$
Breadth-wise parallel `TestSet_C`	$0.063s$	$2.107s$
`FullCTP_C`	$0.035s$	$3.980s$

Table 2.8: Computation times for the comparisons performed in Sections 2.8.1 and 2.8.3. (a): Median time required to compute the TMTI statistic for different n. (b): Median time required to adjust a single p-value without early stopping. (c): Median time required to adjust 10 p-values and detect that the 11^{th} adjusted p-value is not significant using early stopping. (d): Median time required to adjust all p-values. All units are in seconds. The local test is a Bonferroni correction in (b)-(d). In (c), the depth-wise parallelization uses six cores and chunksize 5,000 and the breadth-wise parallelization uses six cores and chunksize six. In (d), the parallelization uses six cores and chunksize $\lceil m/6 \rceil$. All computations were performed on an Apple M1 Pro 3.2 GHz CPU. All times are based on 100 replicate calls.

Chapter 3

Causal discovery in time series

In this chapter, we present the paper 'Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values'. This paper arose from the *Causality 4 Climate* (C4C) competition (Runge et al., 2019) at the 2019 Conference on Neural Information Processing Systems (NeurIPS). The task of this competition was to learn summary graphs using various weather and climate-related time series datasets. The competition consisted of 34 different datasets. For each data set, the participants had to upload a score matrix for the (unknown) summary graph of the underlying time series. The team from the University of Copenhagen, consisting of the paper's authors, finished first on 18 of 34 datasets and second on the remaining 16 datasets. Our work in this competition invited a paper, presented in Paper **B**, which was accepted at NeurIPS 2019 and accompanied by a talk.[1]

In Section 3.2 of this thesis, we discuss the scoring metric used in the C4C competition, and in Section 3.3 we discuss related work published after Paper **B**.

[1] Accessible at https://slideslive.com/38922875, at the time of writing

3.1 Paper B

Sebastian Weichwald, Martin E. Jakobsen, Phillip B. Mogensen, Lasse Petersen, Nikolaj Thams, Gherardo Varando. 'Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values' Proceedings of the NeurIPS 2019 Competition and Demonstration Track, PMLR 123:27-36, 2020.

(Preprint: Weichwald, Sebastian, et al. "Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values." arXiv preprint arXiv:2002.09573 (2020).)

Proceedings of Machine Learning Research 123:27–36, 2020 NeurIPS 2019 Competition and Demonstration Track

Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values

Sebastian Weichwald SWEICHWALD@MATH.KU.DK
Martin E Jakobsen M.JAKOBSEN@MATH.KU.DK
Phillip B Mogensen PBM@MATH.KU.DK
Lasse Petersen LP@MATH.KU.DK
Nikolaj Thams THAMS@MATH.KU.DK
Gherardo Varando GHERARDO.VARANDO@MATH.KU.DK
Copenhagen Causality Lab, Department of Mathematical Sciences, University of Copenhagen

Editors: Hugo Jair Escalante and Raia Hadsell

Abstract

In this article, we describe the algorithms for causal structure learning from time series data that won the Causality 4 Climate competition at the Conference on Neural Information Processing Systems 2019 (NeurIPS). We examine how our combination of established ideas achieves competitive performance on semi-realistic and realistic time series data exhibiting common challenges in real-world Earth sciences data. In particular, we discuss a) a rationale for leveraging linear methods to identify causal links in non-linear systems, b) a simulation-backed explanation as to why large regression coefficients may predict causal links better in practice than small p-values and thus why normalising the data may sometimes hinder causal structure learning. For benchmark usage, we detail the algorithms here and provide implementations at github.com/sweichwald/tidybench. We propose the presented competition-proven methods for baseline benchmark comparisons to guide the development of novel algorithms for structure learning from time series.

Keywords: Causal discovery, structure learning, time series, scaling.

1. Introduction

Inferring causal relationships from large-scale observational studies is an essential aspect of modern climate science (Runge et al., 2019a,b). However, randomised studies and controlled interventions cannot be carried out, due to both ethical and practical reasons. Instead, simulation studies based on climate models are state-of-the-art to study the complex patterns present in Earth climate systems (IPCC, 2013).

Causal inference methodology can integrate and validate current climate models and can be used to probe cause-effect relationships between observed variables. The Causality 4 Climate (C4C) NeurIPS competition (Runge et al., 2020) aimed to further the understanding and development of methods for structure learning from time series data exhibiting common challenges in and properties of realistic weather and climate data.

Structure of this work Section 2 introduces the structure learning task considered. In Section 3, we describe our winning algorithms. With a combination of established ideas, our algorithms achieved competitive performance on semi-realistic data across all 34 challenges in the C4C competition track. Furthermore, at the time of writing, our algorithms lead the rankings for all hybrid and realistic data set categories available on the CauseMe.net benchmark platform which also offers additional synthetic data categories (Runge et al., 2019a). These algorithms—which can be implemented in a few lines of code—are built on simple methods, are computationally efficient, and exhibit solid performance across a variety of different data sets. We therefore encourage the use of these algorithms as baseline benchmarks and guidance of future algorithmic and methodological developments for structure learning from time series.

Beyond the description of our algorithms, we aim at providing intuition that can explain the phenomena we have observed throughout solving the competition task. First, if we *only* ask whether a causal link exists in some non-linear time series system, then we may sidestep the extra complexity of explicit non-linear model extensions (cf. Section 4). Second, when data has a meaningful natural scale, it may—somewhat unexpectedly—be advisable to forego data normalisation and to use raw (vector auto)-regression coefficients instead of p-values to assess whether a causal link exists or not (cf. Section 5).

2. Causal structure learning from time-discrete observations

The task of inferring the causal structure from observational data is often referred to as 'causal discovery' and was pioneered by Pearl (2009) and Spirtes et al. (2001). Much of the causal inference literature is concerned with structure learning from independent and identically distributed (iid) observations. Here, we briefly review some aspects and common assumptions for causally modelling time-evolving systems. More detailed and comprehensive information can be found in the provided references.

Time-discrete observations We may view the discrete-time observations as arising from an underlying continuous-time causal system (Peters et al., 2020). While difficult to conceptualise, the correspondence between structural causal models and differential equation models can be made formally precise (Mooij et al., 2013; Rubenstein et al., 2018; Bongers and Mooij, 2018). Taken together, this yields some justification for modelling dynamical systems by discrete-time causal models.

Summary graph as inferential target It is common to assume a time-homogeneous causal structure such that the dynamics of the observation vector X are governed by $X^t := F(X^{\mathrm{past}(t)}, N^t)$ where the function F determines the next observation based on past values $X^{\mathrm{past}(t)}$ and the noise innovation N^t. Here, structure learning amounts to identifying the summary graph with adjacency matrix A that summarises the causal structure in the following sense: the $(i,j)^{\mathrm{th}}$ entry of the matrix A is 1 if $X_i^{\mathrm{past}(t)}$ enters the structural equation of X_i^t via the i^{th} component of F and 0 otherwise. If $A_{ij} = 1$, we say that "X_i causes X_j". While summary graphs can capture the existence and non-existence of cause-effect relationships, they do in general not correspond to a time-agnostic structural causal model that admits a causal semantics consistent with the underlying time-resolved structural causal model (Rubenstein et al., 2017; Janzing et al., 2018).

Time structure may be helpful for discovery In contrast to the iid setting, the Markov equivalence class of the summary graph induced by the structural equations of a dynamical system is a singleton when assuming causal sufficiency and no instantaneous effects (Peters et al., 2017; Mogensen and Hansen, 2020). This essentially yields a justification and a constraint-based causal inference perspective on Wiener-Granger-causality (Wiener, 1956; Granger, 1969; Peters et al., 2017).

Challenges for causal structure learning from time series data Structure learning from time series is a challenging task hurdled by further problems such as time-aggregation, time-delays, and time-subsampling. All these challenges were considered in the C4C competition and are topics of active research (Danks and Plis, 2013; Hyttinen et al., 2016).

3. The time series discovery benchmark (`tidybench`): Winning algorithms

We developed four simple algorithms,

`SLARAC`	Subsampled Linear Auto-Regression Absolute Coefficients (cf. Alg. 1)
`QRBS`	Quantiles of Ridge regressed Bootstrap Samples (cf. Alg. 2)
`LASAR`	LASso Auto-Regression
`SELVAR`	Selective auto-regressive model

which came in first in 18 and close second in 13 out of the 34 C4C competition categories and won the overall competition (Runge et al., 2020). Here, we provide detailed descriptions of the `SLARAC` and `QRBS` algorithms. Analogous descriptions for the latter two algorithms and implementations of all four algorithms are available at github.com/sweichwald/tidybench.

All of our algorithms output an edge score matrix that contains for each variable pair (X_i, X_j) a score that reflects how likely it is that the edge $X_i \to X_j$ exists. Higher scores correspond to edges that are inferred to be more likely to exist than edges with lower scores, based on the observed data. That is, we rank edges relative to one another but do not perform hypothesis tests for the existence of individual edges. A binary decision can be obtained by choosing a cut-off value for the obtained edge scores. In the C4C competition, submissions were compared to the ground-truth cause-effect adjacency matrix and assessed based on the achieved ROC-AUC when predicting which causal links exist.

The idea behind our algorithms is the following: regress present on past values and inspect the regression coefficients to decide whether one variable is a Granger-cause of another. `SLARAC` fits a VAR model on bootstrap samples of the data each time choosing a random number of lags to include; `QRBS` considers bootstrap samples of the data and Ridge-regresses time-deltas $X(t) - X(t-1)$ on the preceding values $X(t-1)$; `LASAR` considers bootstrap samples of the data and iteratively—up to a maximum lag—LASSO-regresses the residuals of the preceding step onto values one step further in the past and keeps track of the variable selection at each lag to fit an OLS regression in the end with only the selected variables at selected lags included; and `SELVAR` selects edges employing a hill-climbing procedure based on the leave-one-out residual sum of squares and finally scores the selected edges with the absolute values of the regression coefficients. In the absence of instantaneous effects and hidden confounders, Granger-causes are equivalent to a variable's causal parents (Peters et al., 2017, Theorem 10.3). In Section 5, we argue that the size of

the regression coefficients may in certain scenarios be more informative about the existence of a causal link than standard test statistics for the hypothesis of a coefficient being zero. It is argued that for additive noise models, information about the causal ordering may be contained in the raw marginal variances. In test statistics such as the F- and T-statistics, this information is lost when normalising by the marginal variances.

4. Capturing non-linear cause-effect links by linear methods

We explain the rationale behind our graph reconstruction algorithms and how they may capture non-linear dynamics despite being based on linearly regressing present on past values. For simplicity we will outline the idea in a multivariate regression setting with additive noise, but it extends to the time series setting by assuming time homogeneity.

Let $N, X(t_1), X(t_2) \in \mathbb{R}^d$ be random variables such that $X(t_2) := F(X(t_1)) + N$ for some differentiable function $F = (F_1, \ldots, F_d) : \mathbb{R}^d \to \mathbb{R}^d$. Assume that N has mean zero, that it is independent from $X(t_1)$, and that it has mutually independent components. For each $i, j = 1, \ldots, d$ we define the quantity of interest

$$\theta_{ij} = \mathbb{E}\left|\partial_i F_j\left(X(t_1)\right)\right|,$$

such that θ_{ij} measures the expected effect from $X_i(t_1)$ to $X_j(t_2)$. We take the matrix $\Theta = \left(\mathbf{1}_{\theta_{ij}>0}\right)$ as the adjacency matrix of the summary graph between $X(t_1)$ and $X(t_2)$.

In order to detect regions with non-zero gradients of F we create bootstrap samples $\mathcal{D}_1, \ldots, \mathcal{D}_B$. On each bootstrap sample $\mathcal{D}_b$ we obtain the regression coefficients $\widehat{A}_b$ as estimate of the directional derivatives by a (possibly penalised) linear regression technique. Intuitively, if θ_{ij} were zero, then on any bootstrap sample we would obtain a small non-zero contribution. Conversely, if θ_{ij} were non-zero, then we may for some bootstrap samples obtain a linear fit of $X_j(t_2)$ with large absolute regression coefficient for $X_i(t_1)$. The values obtained on each bootstrap sample are then aggregated by, for example, taking the average of the absolute regression coefficients $\widehat{\theta}_{ij} = \frac{1}{B}\sum_{b=1}^{B}\left|(\widehat{A}_b)_{ij}\right|$.

This amounts to searching the predictor space for an effect from $X_i(t_1)$ to $X_j(t_2)$, which is approximated linearly. It is important to aggregate the absolute values of the coefficients to avoid cancellation of positive and negative coefficients. The score $\widehat{\theta}_{ij}$ as such contains no information about whether the effect from $X_i(t_1)$ to $X_j(t_2)$ is positive or negative and it cannot be used to predict $X_j(t_2)$ from $X_i(t_1)$. It serves as a score for the existence of a link between the two variables. This rationale explains how linear methods may be employed for edge detection in non-linear settings without requiring extensions of Granger-type methods that explicitly model the non-linear dynamics and hence come with additional sample complexity (Marinazzo et al., 2008, 2011; Stramaglia et al., 2012, 2014).

5. Large regression coefficients may predict causal links better in practice than small p-values

This section aims at providing intuition behind two phenomena: We observed a considerable drop in the accuracy of our edge predictions whenever 1) we normalised the data or 2) used the T-statistics corresponding to testing the hypothesis of regression coefficients being zero to score edges instead of the coefficients' absolute magnitude. While one could

Algorithm 1: Subsampled Linear Auto-Regression Absolute Coefficients (`SLARAC`)

Input : Data $\boldsymbol{X}$ with T time samples $\boldsymbol{X}(1), \ldots, \boldsymbol{X}(T)$ over d variables.
Parameters : Max number of lags, $L \in \mathbb{N}$.
Number of bootstrap samples, $B \in \mathbb{N}$.
Individual bootstrap sample sizes, $\{v_1, \ldots, v_B\}$.
Output : A $d \times d$ real-valued score matrix, $\widehat{A}$.

Initialise A_{full} as a $d \times dL$ matrix of zeros and $\widehat{A}$ as an empty $d \times d$ matrix;
for $b = 1, \ldots, B$ **do**
  lags $\leftarrow$ random integer in $\{1, \ldots, L\}$;
  Draw a bootstrap sample $\{t_1, \ldots, t_{v_b}\}$ from $\{\text{lags}+1, \ldots, T\}$ with replacement;
  $\boldsymbol{Y}^{(b)} \leftarrow (\boldsymbol{X}(t_1), \ldots \boldsymbol{X}(t_{v_b}))$;
  $\boldsymbol{X}^{(b)}_{\text{past}} \leftarrow \begin{pmatrix} \boldsymbol{X}(t_1 - 1) & \cdots & \boldsymbol{X}(t_1 - \text{lags}) \\ \vdots & \ddots & \vdots \\ \boldsymbol{X}(t_{v_b} - 1) & \cdots & \boldsymbol{X}(t_{v_b} - \text{lags}) \end{pmatrix}$;
  Fit OLS estimate β of regressing $\boldsymbol{Y}^{(b)}$ onto $\boldsymbol{X}^{(b)}_{\text{past}}$;
  Zero-pad β such that $\dim \beta = d \times dL$;
  $A_{\text{full}} \leftarrow A_{\text{full}} + |\beta|$;
end
Aggregate $(\widehat{A})_{i,j} \leftarrow \max((A_{\text{full}})_{i,j+0\cdot d}, \ldots, (A_{\text{full}})_{i,j+L\cdot d})$ for every i, j;
Return: Score matrix $\widehat{A}$.

Algorithm 2: Quantiles of Ridge regressed Bootstrap Samples (`QRBS`)

Input : Data $\boldsymbol{X}$ with T time samples $\boldsymbol{X}(1), \ldots, \boldsymbol{X}(T)$ over d variables.
Parameters : Number of bootstrap samples, $B \in \mathbb{N}$.
Size of bootstrap samples, $v \in \mathbb{N}$.
Ridge regression penalty, $\kappa \geq 0$.
Quantile for aggregating scores, $q \in [0, 1]$.
Output : A $d \times d$ real-valued score matrix, $\widehat{A}$.

for $b = 1, \ldots, B$ **do**
  Draw a bootstrap sample $\{t_1, \ldots, t_v\}$ from $\{2, \ldots, T\}$ with replacement;
  $\boldsymbol{Y}^{(b)} \leftarrow (\boldsymbol{X}(t_1) - \boldsymbol{X}(t_1 - 1), \ldots, \boldsymbol{X}(t_v) - \boldsymbol{X}(t_v - 1))$;
  $\boldsymbol{X}^{(b)} \leftarrow (\boldsymbol{X}(t_1 - 1), \ldots, \boldsymbol{X}(t_v - 1))$;
  Fit a ridge regression of $\boldsymbol{Y}^{(b)}$ onto $\boldsymbol{X}^{(b)}$: $\widehat{A}_b = \arg\min_A \|\boldsymbol{Y}^{(b)} - A\boldsymbol{X}^{(b)}\| + \kappa\|A\|$;
end
Aggregate $\widehat{A} \leftarrow q^{th}$ element-wise quantile of $\{|\widehat{A}_1|, \ldots, |\widehat{A}_B|\}$;
Return Score matrix $\widehat{A}$.

try to attribute these phenomena to some undesired artefact in the competition setup, it is instructive to instead try to understand when exactly one would expect such behaviour.

We illustrate a possible explanation behind these phenomena and do so in an iid setting in favour of a clear exposition, while the intuition extends to settings of time series observations and our proposed algorithms. The key remark is, that under comparable noise variances, the variables' marginal variances tend to increase along the causal ordering. If data are observed at comparable scales—say sea level pressure in different locations measured in the same units—or at scales that are in some sense naturally relative to the true data generating mechanism, then absolute regression coefficients may be preferable to T-test statistics. Effect variables tend to have larger marginal variance than their causal ancestors. This helpful signal in the data is diminished by normalising the data or the rescaling when computing the T-statistics corresponding to testing the regression coefficients for being zero. This rationale is closely linked to the identifiability of Gaussian structural equation models under equal error variances Peters and Bühlmann (2014). Without any prior knowledge about what physical quantities the variables correspond to and their natural scales, normalisation remains a reasonable first step. We are not advocating that one should use the raw coefficients and not normalise data, but these are two possible alterations of existing structure learning procedures that may or may not, depending on the concrete application at hand, be worthwhile exploring. Our algorithms do not perform data normalisation, so the choice is up to the user whether to feed normalised or raw data, and one could easily change to using p-values or T-statistics instead of raw coefficients for edge scoring.

5.1. Instructive iid case simulation illustrates scaling effects

We consider data simulated from a standard acyclic linear Gaussian model. Let $N \sim \mathcal{N}\left(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)\right)$ be a d-dimensional random variable and let $\boldsymbol{B}$ be a $d \times d$ strictly lower-triangular matrix. Further, let X be a d-valued random variable constructed according to the structural equation $X = \boldsymbol{B}X + N$, which induces a distribution over X via $X = (I - \boldsymbol{B})^{-1}N$. We have assumed, without loss of generality, that the causal order is aligned such that X_i is further up in the causal order than X_j whenever $i < j$. We ran 100 repetitions of the experiment, each time sampling a random lower triangular 50×50-matrix $\boldsymbol{B}$ where each entry in the lower triangle is drawn from a standard Gaussian with probability $1/4$ and set to zero otherwise. For each such obtained $\boldsymbol{B}$ we sample $n = 200$ observations from $X = \boldsymbol{B}X + N$ which we arrange in a data matrix $\boldsymbol{X} \in \mathbb{R}^{200\times 50}$ of zero-centred columns denoted by $\boldsymbol{X}_j$.

We regress each X_j onto all remaining variables $X_{\neg j}$ and compare scoring edges $X_i \to X_j$ by the absolute values of a) the regression coefficients $|\widehat{b}_{i\to j}|$, versus b) the T-statistics $|\widehat{t}_{i\to j}|$ corresponding to testing the hypothesis that the regression coefficient $\widehat{b}_{i\to j}$ is zero. That is, we consider

$$|\widehat{b}_{i\to j}| = \left|(\boldsymbol{X}_{\neg j}^\top \boldsymbol{X}_{\neg j})^{-1}\boldsymbol{X}_{\neg j}^\top \boldsymbol{X}_j\right|_i$$

versus

$$|\widehat{t}_{i\to j}| = |\widehat{b}_{i\to j}|\sqrt{\frac{\widehat{\mathrm{var}}(X_i|X_{\neg i})}{\widehat{\mathrm{var}}(X_j|X_{\neg j})}}\sqrt{\frac{(n-d)}{\left(1 - \widehat{\mathrm{corr}}^2(X_i, X_j|X_{\neg\{i,j\}})\right)}} \tag{1}$$

where $\widehat{\text{var}}(X_j|X_{\neg j})$ is the residual variance after regressing X_j onto the other variables $X_{\neg j}$, and $\widehat{\text{corr}}(X_i, X_j|X_{\neg\{i,j\}})$ is the residual correlation between X_i and X_j after regressing both onto the remaining variables.

We now compare, across three settings, the AUC obtained by either using the absolute value of the regression coefficients $|\widehat{b}_{i\to j}|$ or the absolute value of the corresponding T-statistics $|\widehat{t}_{i\to j}|$ for edge scoring. Results are shown in the left, middle, and right panel of Figure 1, respectively.

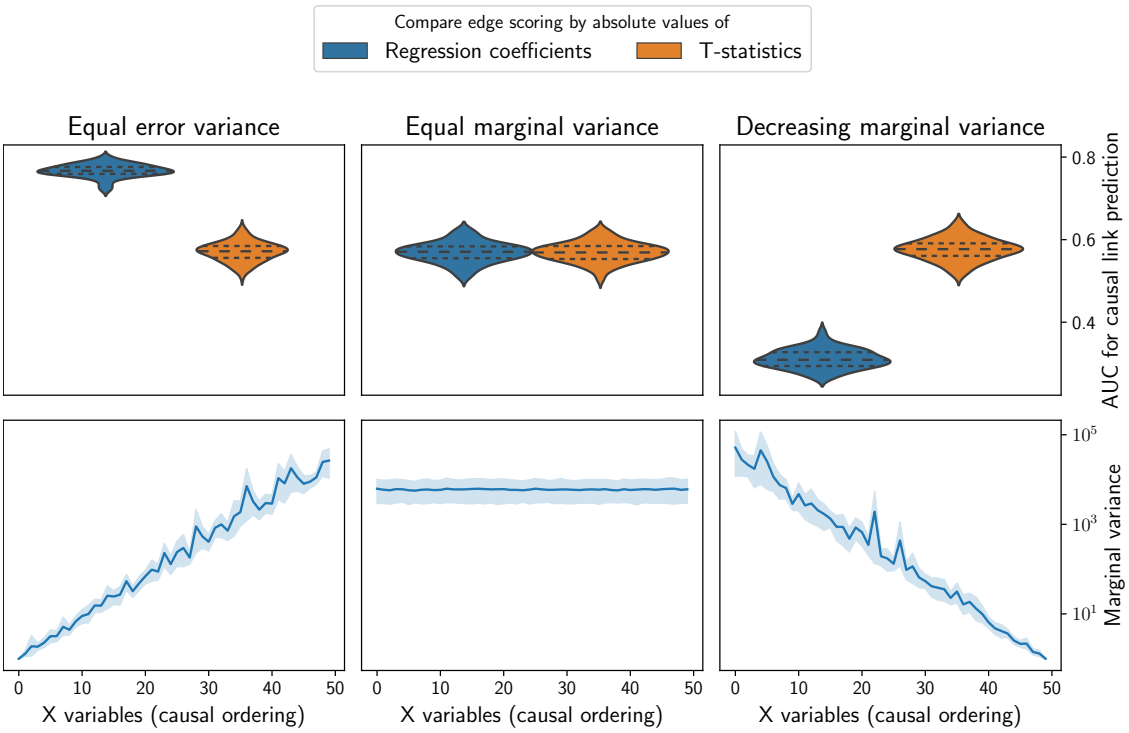


Figure 1: Results of the simulation experiment described in Section 5.1. Data is generated from an acyclic linear Gaussian model, in turn each variable is regressed onto all remaining variables and either the raw regression coefficient $|\widehat{b}_{i\to j}|$ or the corresponding T-statistics $|\widehat{t}_{i\to j}|$ is used to score the existence of an edge $i \to j$. The top row shows the obtained AUC for causal link prediction and the bottom row the marginal variance of the variables along the causal ordering. The left panel shows naturally increasing marginal variance for equal error variances, for the middle and right panel the model parameters and error variances are rescaled to enforce equal and decreasing marginal variance, respectively.

In the setting with equal error variances $\sigma_i^2 = \sigma_j^2 \ \forall i, j$**,** we observe that i) the absolute regression coefficients beat the T-statistics for edge predictions in terms of AUC, and ii) the marginal variances naturally turn out to increase along the causal ordering.

When moving from $|\widehat{b}_{i\to j}|$ to $|\widehat{t}_{i\to j}|$ for scoring edges, we multiply by a term that compares the relative residual variance of X_i and X_j. If X_i is before X_j in the causal ordering it tends to have both smaller marginal and—in our simulation set-up—residual variance than X_j as it becomes increasingly more difficult to predict variables further down the causal ordering. In this case, the fraction of residual variances will tend to be smaller than one and consequently the raw regression coefficients $|\widehat{b}_{i\to j}|$ will be shrunk when moving to $|\widehat{t}_{i\to j}|$. This can explain the worse performance of the T-statistics compared to the raw regression coefficients for edge scoring as scores will tend to be shrunk when in fact $X_i \to X_j$.

Enforcing equal marginal variances by rescaling the rows of B and the σ_i^2's, we indeed observe that regression coefficients and T-statistics achieve comparable performance in edge prediction in this somewhat artificial scenario. Here, neither the marginal variances nor the residual variances appear to contain information about the causal ordering any more and the relative ordering between regression coefficients and T-statistics is preserved when multiplying by the factor highlighted in Equation 1.

Enforcing decreasing marginal variances by rescaling the rows of B and the σ_i^2's, we can, in line with our above reasoning, indeed obtain an artificial scenario in which the T-statistics will outperform the regression coefficients in edge prediction, as now, the factors we multiply by will work in favour of the T-statistics.

6. Conclusion and Future Work

We believe competitions like the Causality 4 Climate competition (Runge et al., 2020) and causal discovery benchmark platforms like CauseMe.net (Runge et al., 2019a) are important for bundling and informing the community's joint research efforts into methodology that is readily applicable to tackle real-world data. In practice, there are fundamental limitations to causal structure learning that ultimately require us to employ untestable causal assumptions to proceed towards applications at all. Yet, both these limitations and assumptions are increasingly well understood and characterised by methodological research and time and again need to be challenged and examined through the application to real-world data.

Beyond the algorithms presented here and proposed for baseline benchmarks, different methodology as well as different benchmarks may be of interest. For example, our methods detect causal links and are viable benchmarks for the structure learning task but they do not per se enable predictions about the interventional distributions.

Acknowledgments

The authors thank Niels Richard Hansen, Steffen Lauritzen, and Jonas Peters for insightful discussions. Thanks to the organisers for a challenging and insightful Causality 4 Climate NeurIPS competition. NT was supported by a research grant (18968) from VILLUM FONDEN. LP and GV were supported by a research grant (13358) from VILLUM FONDEN. MEJ and SW were supported by the Carlsberg Foundation.

References

S. Bongers and J. M. Mooij. From random differential equations to structural causal models: The stochastic case. *arXiv preprint arXiv:1803.08784*, 2018.

D. Danks and S. Plis. Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings*, 2013.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

A. Hyttinen, S. Plis, M. Järvisalo, F. Eberhardt, and D. Danks. Causal Discovery from Subsampled Time Series Data by Constraint Optimization. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, 2016.

IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2013.

D. Janzing, P. K. Rubenstein, and B. Schölkopf. Structural causal models for macro-variables in time-series. *arXiv preprint arXiv:1804.03911*, 2018.

D. Marinazzo, M. Pellicoro, and S. Stramaglia. Kernel-Granger causality and the analysis of dynamical networks. *Physical Review E*, 77(5):056215, 2008.

D. Marinazzo, W. Liao, H. Chen, and S. Stramaglia. Nonlinear connectivity by Granger causality. *NeuroImage*, 58(2):330 – 338, 2011.

S. W. Mogensen and N. R. Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1):539–559, 2020.

J. M. Mooij, D. Janzing, and B. Schölkopf. From Ordinary Differential Equations to Structural Causal Models: the deterministic case. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2013.

J. Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

J. Peters, S. Bauer, and N. Pfister. Causal models for dynamical systems. *arXiv preprint arXiv:2001.06208*, 2020.

P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2017.

P. K. Rubenstein, S. Bongers, J. M. Mooij, and B. Schölkopf. From deterministic ODEs to dynamic structural causal models. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2018.

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, E. H. Muñoz-Marí, J. andand van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1):2553, 2019a.

J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11), 2019b.

J. Runge, X.-A. Tibau, M. Bruhns, J. Muñoz-Marí, and G. Camps-Valls. The causality for climate competition. In Hugo Jair Escalante and Raia Hadsell, editors, *PMLR NeurIPS Competition & Demonstration Track Postproceedings*, Proceedings of Machine Learning Research. PMLR, 2020. URL https://causeme.uv.es/. Forthcoming.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2 edition, 2001.

S. Stramaglia, G.-R. Wu, M. Pellicoro, and D. Marinazzo. Expanding the transfer entropy to identify information circuits in complex systems. *Physical Review E*, 86(6):066211, 2012.

S. Stramaglia, J. M. Cortes, and D. Marinazzo. Synergy and redundancy in the Granger causal analysis of dynamical networks. *New Journal of Physics*, 16(10):105003, 2014.

N. Wiener. The theory of prediction. *Modern Mathematics for Engineers*, 1956.

3.2 A discussion of performance metrics for causal discovery methods

In this section, we discuss the performance metric used in the *Causality 4 Climate* competition (Runge et al., 2019).

The task of the C4C competition was to estimate the $d \times d$ adjacency matrix A of the summary graphs (defined in Section 1.3.6) of different time series $X(t)$. These time series consisted of a mixture of simulated and real-world data. In all cases, the ground truth summary graph was known to the competition organizers but not the participants. Note that in the context of the C4C competition, we enforce diag(A) = 1 (and similarly for any estimate), meaning that all marginal variables are affected by their pasts.

In the competition, participants had to estimate A by supplying a normalized score matrix to an online platform.[2] A score matrix is a matrix where each element (i, j) represents our confidence in the existence of a causal link $(i \to j)$. Thus, a score matrix is a natural estimate to present. However, many real-world applications require that one makes binary decisions about the existence of causal links. The score matrix can be converted to a binary estimate $\hat{A}(\tau)$ of A by thresholding the elements by some parameter τ, i.e.,

$$\hat{A}_{i,j}(\tau) = \begin{cases} 1, & \text{score for edge } (i \to j) \geq \tau \\ 0, & \text{else} \end{cases}.$$

If we had specified a threshold τ_0 a priori, the performance of the algorithms we developed could be evaluated, e.g., by the Structural Hamming Distance (SHD) or Structural Intervention Distance (SID) (Peters and Bühlmann, 2015).[3] However, there is no immediate, systematic way of choosing the threshold. Thus, the output of our algorithms depends directly on a user-specific threshold. This critique applies to many structure-learning algorithms, however. For example, in many constraint-based structure learning algorithms, one determines the existence of a causal link by a conditional independence test, which is thresholded at a user-chosen significance level. While each marginal conditional independence test has the interpretation of controlling the Type I error, this interpretation disappears when we perform many sequential tests. Thus, the significance level is also a user-specific thresholding parameter.

Still, estimating a non-binary score matrix may be useful for exploratory purposes. To evaluate the performance of our algorithms, we let $\mathrm{FPR}(\tau)$ and $\mathrm{TPR}(\tau)$ denote the False Positive Rate and True Positive Rate of the score matrix, respectively. That is,

$$\mathrm{FPR}(\tau) = \frac{\#\text{non-edges in } \mathcal{G} \text{ with score} \geq \tau}{\#\text{non-edges in } \mathcal{G}}$$

and

$$\mathrm{TPR}(\tau) = \frac{\#\text{edges in } \mathcal{G} \text{ with score} \geq \tau}{\#\text{edges in } \mathcal{G}},$$

[2] https://causeme.uv.es

[3] The SHD (resp. SID) counts the number of incorrectly inferred edges (resp. interventional distributions) in our estimate $\hat{A}(\tau_0)$ of A.

for $\tau \in (0, 1)$. In the C4C competition, the performance of our algorithms was assessed using the Area under the Receiver Operating Characteristic curve (ROC-AUC), defined as the area under the curve $\tau \mapsto (\text{FPR}(\tau), \text{TPR}(\tau))$ (Fawcett, 2006). In practice, we do not observe a curve, but rather a finite set of points that we connect and integrate using the trapezoidal rule, while forcing the connected line to pass through the points $(0, 0)$ and $(1, 1)$. It is common in machine learning to employ ROC-AUC as a performance metric. An upside of this metric is that one does not need to perform any thresholding, and we may instead keep the non-binary score matrix. Then, if the score matrix has a high ROC-AUC, the interpretation is that edges (resp. non-edges) in $\mathcal{G}$ have a high (resp. low) associated edge score, on average.

Below, we give an – admittedly very artificial – example to illustrate that it is possible to obtain a ROC-AUC close to one while inferring a diverging number of incorrect interventional distributions. This example is *not* meant to showcase that the ROC-AUC is an inherently poor performance metric, but rather that causal discovery is a difficult task and that performing well by one single metric does not necessarily mean that one has fully learned the causal system under investigation.

Example 3.1. *Consider the two DAGs, $\mathcal{G}_1$ and $\mathcal{G}_2$, shown in Figure 3.1. Sup-*

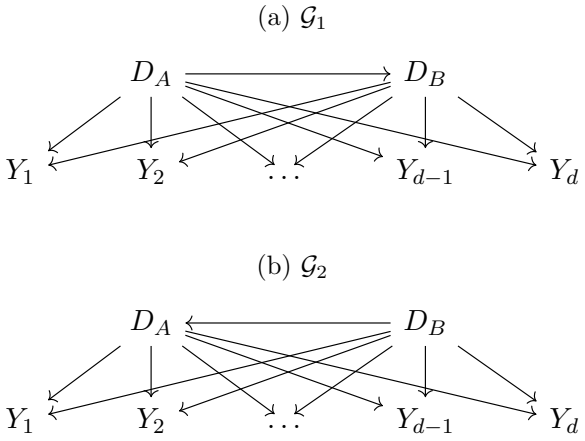


Figure 3.1: (a): The true, data-generating graph $\mathcal{G}_1$. (b): The estimated graph $\mathcal{G}_2$.

pose that we observe data generated by the distribution implied by $\mathcal{G}_1$. Using the observed data, we have estimated a score matrix, and find that we:

1. *Correctly estimate the score for all edges $D_A \to Y_i$ and $D_B \to Y_i$ to be exactly one.*

2. *Incorrectly estimate the scores for the edges $D_B \to D_A$ and $D_A \to D_B$ to be exactly one and zero, respectively.*

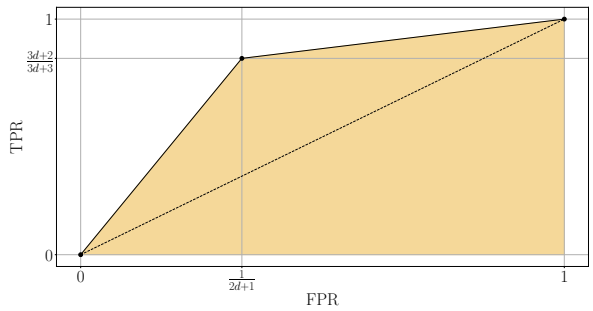


Figure 3.2: Observed ROC curve in Example 3.1. The observed ROC-AUC, indicated in yellow, is $(12d^2 + 13d + 2)/(12d^2 + 18d + 6)$.

3. Correctly estimate the scores of the remaining non-edges to be exactly 0.

In other words, no matter the choice of $\tau \in (0, 1)$*, we have estimated* $\mathcal{G}_2$*. In this case, we have for all* $\tau \in (0, 1)$

$$\mathrm{TPR}(\tau) = \frac{3d+2}{3d+3},^{4} \quad \textit{and} \quad \mathrm{FPR}(\tau) = \frac{1}{2d+1}.$$

Thus, we observe $\mathrm{FPR} = (0, (2d+1)^{-1}, 1)$ *and* $\mathrm{TPR} = (0, (3d+2)/(3d+3), 1)$*. The observed Receiver Operating Characteristic (ROC) curve is shown in Figure 3.2. Integrating the ROC curve using the trapezoidal rule we find, with a small rewrite, that*

$$\text{ROC-AUC}_d = \frac{12d^2 + 13d + 2}{12d^2 + 18d + 6}.$$

Clearly, we have $\lim_{d\to\infty} \text{ROC-AUC}_d = 1$*. Thus, as the size of the system grows, the ROC-AUC goes to one.*

In this case, where we do not perform any thresholding, we can measure the closeness between $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *by the SHD or SID directly. The SHD between the estimated and true graphs is constant at one.*[5] *However, the SID between* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *is* $2d+2$*. In particular, we incorrectly infer the following interventional distributions for any* $i \in [d]$*:*

$$\mathbb{P}_{D_A}^{\mathrm{do}(D_B=c)}, \quad \mathbb{P}_{D_B}^{\mathrm{do}(D_A=c)}, \quad \mathbb{P}_{Y_i}^{\mathrm{do}(D_B=c)}, \quad \textit{and} \quad \mathbb{P}_{Y_i}^{\mathrm{do}(D_A=c)}.$$

Not only does the number of incorrectly inferred distributions diverge, but the estimated distributions can also be arbitrarily wrong under hard interventions. For instance, if all assignments are linear with coefficient one and all noise innovations are i.i.d. $N(0, 1)$*-distributed, then* $\mathbb{P}_{Y_i}^{\mathrm{do}(D_B)=c} = N(c, 2)$ *in* $\mathcal{G}_1$*, but* $\mathbb{P}_{Y_i}^{\mathrm{do}(D_B)=c} = N(2c, 2)$ *in* $\mathcal{G}_2$*.*

[4] Recall that we enforce $\mathrm{diag}(A) = (1, \ldots, 1)$, meaning that the $2d$ edges from D_A, D_B to all Y_i's are correctly inferred, as well as the $d + 2$ self-cycles in $\mathcal{G}$.

[5] Or two, depending on whether we count the misdirection of an edge as one or two mistakes.

3.3 Simulation artifacts of Additive Noise Models

In Paper **B**, we argued that – in some scenarios – structure learning algorithms may perform better when using the magnitude of regression coefficients as evidence of causal links rather than p-values. This fact may appear counter-intuitive, as the magnitude of a regression coefficient depends on the scale of the input predictor. The rationale behind using regression coefficients as scores for the existence of causal links in the competition was that the algorithms were, to some degree, tailored to the problem at hand. Here, we observed that the performance of our algorithms dropped when using p-values instead of regression coefficients or when normalizing data prior to regression. We argued in Paper **B**, Section 5 that this drop in performance was likely because the ordering of marginal variances approximately equaled the causal order of the input variables. However, we do not generally have evidence that real-world data exhibits this tendency of increasing marginal variances. Thus, we do not advocate using regression coefficient magnitudes over p-values in general, but rather that this approach can serve as a benchmark.

After the publication of Paper **B**, the phenomenon of increasing marginal variances has been further studied by Reisach et al. (2021), who coin the term 'varsortability'. In this paper, the authors show that varsortability is a common phenomenon of simulated Additive Noise Models (ANMs),[6] and that the causal ordering of an ANM is partially identifiable from its marginal variances. Furthermore, the authors show that continuous structure-learning algorithms (such as GOLEM (Ng et al., 2020) and NOTEARS (Zheng et al., 2018)) and some combinatorial structure-learning algorithms[7] depend on the scale of data. Furthermore, the authors show that the performance of these structure learning algorithms drops when standardizing data prior to analysis. Moreover, the authors show that standardization may not be sufficient by itself; a high degree of varsortability leaves a distinct covariance pattern that cannot be removed by standardization alone. Seng et al. (2022) further shows that the output graph of NOTEARS can be controlled by rescaling individual variables in data before analysis.

Taken together, these papers indicate a need for caution when evaluating the performance of structure learning algorithms on simulated data, as high performance may simply be due to simulation artifacts that may not exist in real-world data. However, if we analyze data in which all variables are measured on the same scale, or in some way have a meaningful natural scale, we can exploit these marginal variance patterns to learn the underlying causal structure.

[6] An ANM is an SCM where each structural assignment takes the form $X_v := N_v + \sum_{k \in \mathrm{PA}_v} f_k(X_k)$.

[7] In particular, combinatorial structure-learning algorithms that are not score equivalent. A score function is score equivalent if all DAGs in the same Markov equivalence class are assigned equal scores.

Chapter 4

Causal discovery in heterogeneous data

This chapter contains Paper **C**: 'Invariant Ancestry Search'. In this paper, we propose a novel method – called Invariant Ancestry Search (IAS) – for causal structure learning. We show that IAS can be used to learn invariant subsets of ancestors of a given response variable when applied to data sampled from heterogeneous environments. Paper **C** was accepted into, and presented at, the thirty-ninth International Conference on Machine Learning (ICML) 2022.[1]

In Section 4.2 of this thesis, we show how the output of IAS can be post hoc separated into parents and non-parental ancestors of Y with high probability.

4.1 Paper C

Phillip B. Mogensen, Nikolaj Thams, Jonas Peters. 'Invariant Ancestry Search' Proceedings of the 39th International Conference on Machine Learning, PMLR 162:15832-15857, 2022.

(Preprint: Mogensen, Phillip B., Nikolaj Thams, and Jonas Peters. 'Invariant Ancestry Search'. arXiv preprint arXiv:2202.00913 (2022).)

[1] At the time of writing, the talk and accompanying poster are available online at `https://icml.cc/virtual/2022/spotlight/16248` and `https://icml.cc/media/PosterPDFs/ICML%202022/569ff987c643b4bedf504efda8f786c2.png`, respectively.

Invariant Ancestry Search

Phillip B. Mogensen [1] **Nikolaj Thams** [1] **Jonas Peters** [1]

Abstract

Recently, methods have been proposed that exploit the invariance of prediction models with respect to changing environments to infer subsets of the causal parents of a response variable. If the environments influence only few of the underlying mechanisms, the subset identified by invariant causal prediction (ICP), for example, may be small, or even empty. We introduce the concept of minimal invariance and propose invariant ancestry search (IAS). In its population version, IAS outputs a set which contains only ancestors of the response and is a superset of the output of ICP. When applied to data, corresponding guarantees hold asymptotically if the underlying test for invariance has asymptotic level and power. We develop scalable algorithms and perform experiments on simulated and real data.

1. Introduction

Causal reasoning addresses the challenge of understanding why systems behave the way they do and what happens if we actively intervene. Such mechanistic understanding is inherent to human cognition, and developing statistical methodology that learns and utilizes causal relations is a key step in improving both narrow and broad AI (Jordan, 2019; Pearl, 2018). Several approaches exist for learning causal structures from observational data. Approaches such as the PC-algorithm (Spirtes et al., 2000) or greedy equivalence search (Chickering, 2002) learn (Markov equivalent) graphical representations of the causal structure (Lauritzen, 1996). Other approaches learn the graphical structure under additional assumptions, such as non-Gaussianity (Shimizu et al., 2006) or non-linearity (Hoyer et al., 2009; Peters et al., 2014). Zheng et al. (2018) convert the problem into a continuous optimization problem, at the expense of identifiability guarantees.

[1]Department of Mathematical Sciences, University of Copenhagen, Denmark. Correspondence to: Phillip B. Mogensen <pbm@math.ku.dk>, Jonas Peters <jonas.peters@math.ku.dk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Invariant causal prediction (ICP) (Peters et al., 2016; Heinze-Deml et al., 2018; Pfister et al., 2019; Gamella & Heinze-Deml, 2020; Martinet et al., 2021) assumes that data are sampled from heterogeneous environments (which can be discrete, categorical or continuous), and identifies direct causes of a target Y, also known as causal parents of Y. Learning ancestors (or parents) of a response Y yields understanding of anticipated changes when intervening in the system. It is a less ambitious task than learning the complete graph but may allow for methods that come with weaker assumptions and stronger guarantees. More concretely, for predictors $X_1, \dots, X_d$, ICP searches for subsets $S \subseteq \{1, \dots, d\}$ that are invariant; a set X_S of predictors is called invariant if it renders Y independent of the environment, conditional on X_S. ICP then outputs the intersection of all invariant predictor sets $S_{\text{ICP}} := \cap_{S \text{ invariant}} S$. Peters et al. (2016) show that if invariance is tested empirically from data at level α, the resulting intersection $\hat{S}_{\text{ICP}}$ is a subset of direct causes of Y with probability at least $1 - \alpha$.[1]

In many cases, however, the set learned by ICP forms a strict subset of all direct causes or may even be empty. This is because disjoint sets of predictors can be invariant, yielding an empty intersection, which may happen both for finite samples as well as in the population setting. In this work, we introduce and characterize minimally invariant sets of predictors, that is, invariant sets S for which no proper subset is invariant. We propose to consider the union S_{IAS} of all minimally invariant sets, where IAS stands for invariant ancestry search. We prove that S_{IAS} is a subset of causal ancestors of Y, invariant, non-empty and contains S_{ICP}. Learning causal ancestors of a response may be desirable for several reasons: e.g., they are the variables that may have an influence on the response variable when intervened on. In addition, because IAS yields an invariant set, it can be used to construct predictions that are stable across environments (e.g., Rojas-Carulla et al., 2018; Christiansen et al., 2022).

In practice, we estimate minimally invariant sets using a test for invariance. If such a test has asymptotic power against some of the non-invariant sets (specified in Section 5.2), we show that, asymptotically, the probability of $\hat{S}_{\text{IAS}}$ being a

[1]Rojas-Carulla et al. (2018); Magliacane et al. (2018); Arjovsky et al. (2019); Christiansen et al. (2022) propose techniques that consider similar invariance statements with a focus on distribution generalization instead of causal discovery.

subset of the ancestors is at least $1-\alpha$. This puts stronger assumptions on the invariance test than ICP (which does not require any power) in return for discovering a larger set of causal ancestors. We prove that our approach retains the ancestral guarantee if we test minimal invariance only among subsets up to a certain size. This yields a computational speed-up compared to testing minimal invariance in all subsets, but comes at the cost of potentially finding fewer causal ancestors.

The remainder of this work is organized as follows. In Section 2 we review relevant background material, and we introduce the concept of minimal invariance in Section 3. Section 4 contains an oracle algorithm for finding minimally invariant sets (and a closed-form expression of S_{ICP}) and Section 5 presents theoretical guarantees when testing minimal invariance from data. In Section 6 we evaluate our method in several simulation studies as well as a real-world data set on gene perturbations. Code is provided at `https://github.com/PhillipMogensen/InvariantAncestrySearch`.

2. Preliminaries

2.1. Structural Causal Models and Graphs

We consider a setting where data are sampled from a structural causal model (SCM) (Pearl, 2009; Bongers et al., 2021)

$$Z_j := f_j(\text{PA}_j, \epsilon_j),$$

for some functions f_j, parent sets PA_j and noise distributions ϵ_j. Following (Peters et al., 2016; Heinze-Deml et al., 2018), we consider an SCM over variables $Z := (E, X, Y)$ where E is an exogenous environment variable (i.e., $\text{PA}_E = \emptyset$), Y is a response variable and $X = (X_1, \dots, X_d)$ is a collection of predictors of Y. We denote by $\mathcal{P}$ the family of all possible distributions induced by an SCM over (E, X, Y) of the above form.

For a collection of nodes $j \in [d] := \{1, \dots, d\}$ and their parent sets PA_j, we define a directed graph $\mathcal{G}$ with nodes $[d]$ and edges $j' \to j$ for all $j' \in \text{PA}_j$. We denote by CH_j, AN_j and DE_j the children, ancestors and descendants of a variable j, respectively, neither containing j. A graph $\mathcal{G}$ is called a directed acyclic graph (DAG) if it does not contain any directed cycles. See Pearl (2009) for more details and the definition of d-separation.

Throughout the remainder of this work, we make the following assumptions about causal sufficiency and exogeneity of E (Section 7 describes how these assumptions can be relaxed).

Assumption 2.1. Data are sampled from an SCM over nodes (E, X, Y), such that the corresponding graph is a DAG, the distribution is faithful with respect to this DAG, and the environments are exogenous, i.e., $\text{PA}_E = \emptyset$.

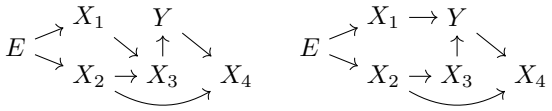


Figure 1. Two structures where $S_{\text{ICP}} \subsetneq \text{PA}_Y$. (*left*) $S_{\text{ICP}} = \emptyset$. (*right*) $S_{\text{ICP}} = \{1\}$. In both, our method outputs $S_{\text{IAS}} = \{1, 2, 3\}$.

2.2. Invariant Causal Prediction

Invariant causal prediction (ICP), introduced by Peters et al. (2016), exploits the existence of heterogeneity in the data, here encoded by an environment variable E, to learn a subset of causal parents of a response variable Y. A subset of predictors $S \subseteq [d]$ is *invariant* if $Y \perp\!\!\!\perp E \mid S$, and we define $\mathcal{I} := \{S \subseteq [d] \mid S \text{ invariant}\}$ to be the set of all invariant sets. We denote the corresponding hypothesis that S is invariant by

$$H_{0,S}^{\mathcal{I}}: \quad S \in \mathcal{I}.$$

Formally, $H_{0,S}^{\mathcal{I}}$ corresponds to a subset of distributions in $\mathcal{P}$, and we denote by $H_{A,S}^{\mathcal{I}} := \mathcal{P} \setminus H_{0,S}^{\mathcal{I}}$ the alternative hypothesis to $H_{0,S}^{\mathcal{I}}$. Peters et al. (2016) define the oracle output

$$S_{\text{ICP}} := \bigcap_{S: H_{0,S}^{\mathcal{I}} \text{ true}} S \tag{1}$$

(with $S_{\text{ICP}} = \emptyset$ if no sets are invariant) and prove $S_{\text{ICP}} \subseteq \text{PA}_Y$. If provided with a test for the hypotheses $H_{0,S}^{\mathcal{I}}$, we can test all sets $S \subseteq [d]$ for invariance and take the intersection over all accepted sets: $\hat{S}_{\text{ICP}} := \bigcap_{S: H_{0,S}^{\mathcal{I}} \text{ not rejected}} S$; If the invariance test has level α, $\hat{S}_{\text{ICP}} \subseteq \text{PA}_Y$ with probability at least $1-\alpha$.

However, even for the oracle output in Equation (1), there are many graphs for which S_{ICP} is a strict subset of PA_Y. For example, in Figure 1 (left), since both $\{1, 2\}$ and $\{3\}$ are invariant, $S_{\text{ICP}} \subseteq \{1, 2\} \cap \{3\} = \emptyset$. This does not violate $S_{\text{ICP}} \subseteq \text{PA}_Y$, but is non-informative. Similarly, in Figure 1 (right), $S_{\text{ICP}} = \{1\}$, as all invariant sets contain $\{1\}$. Here, S_{ICP} contains some information, but is not able to recover the full parental set. In neither of these two cases, S_{ICP} is an invariant set. If the environments are such that each parent of Y is either affected by the environment directly or is a parent of an affected node, then $S_{\text{ICP}} = \text{PA}_Y$ (Peters et al., 2016, proof of Theorem 3). The shortcomings of ICP thus relate to settings where the environments act on too few variables or on uninformative ones.

For large d, it has been suggested to apply ICP to the variables in the *Markov boundary* (Pearl, 2014), $\text{MB}_Y = \text{PA}_Y \cup \text{CH}_Y \cup \text{PA}(\text{CH}_Y)$ (we denote the oracle output by $S_{\text{ICP}}^{\text{MB}}$). As $\text{PA}_Y \subseteq \text{MB}_Y$, it still holds that $S_{\text{ICP}}^{\text{MB}}$ is a sub-

set of the causal parents of the response.[2] However, the procedure must still be applied to $2^{|\,\mathrm{MB}_Y\,|}$ sets, which is only feasible if the Markov boundary is sufficiently small. In practice, the Markov boundary can, for example, be estimated using Lasso regression or gradient boosting techniques (Tibshirani, 1996; Meinshausen & Bühlmann, 2006; Friedman, 2001).

3. Minimal Invariance and Ancestry

We now introduce the concept of minimally invariant sets, which are invariant sets that do not have any invariant subsets. We propose to consider S_{IAS}, the oracle outcome of invariant ancestry search, defined as the union of all minimally invariant sets. We will see that S_{IAS} is an invariant set, it consists only of ancestors of Y, and it contains S_{ICP} as a subset.

Definition 3.1. Let $S \subseteq [d]$. We say that S is *minimally invariant* if and only if

$$S \in \mathcal{I} \text{ and } \forall S' \subsetneq S: \; S' \notin \mathcal{I};$$

that is, S is invariant and no subset of S is invariant. We define $\mathcal{MI} := \{S \mid S \text{ minimally invariant}\}$.

The concept of minimal invariance is closely related to the concept of minimal d-separators (Tian et al., 1998). This connection allows us to state several properties of minimal invariance. For example, an invariant set is minimally invariant if and only if it is non-invariant as soon as one of its elements is removed.

Proposition 3.2. *Let $S \subseteq [d]$. Then $S \in \mathcal{MI}$ if and only if $S \in \mathcal{I}$ and for all $j \in S$, it holds that $S \setminus \{j\} \notin \mathcal{I}$.*

The proof follows directly from (Tian et al., 1998, Corollary 2). We can therefore decide whether a given invariant set S is minimally invariant using $\mathcal{O}(|S|)$ checks for invariance, rather than $\mathcal{O}(2^{|S|})$ (as suggested by Definition 3.1). We use this insight in Section 5.1, when we construct a statistical test for whether or not a set is minimally invariant.

To formally define the oracle outcome of IAS, we denote the hypothesis that a set S is minimally invariant by

$$H_{0,S}^{\mathcal{MI}}: \quad S \in \mathcal{MI}$$

(and the alternative hypothesis, $S \notin \mathcal{MI}$, by $H_{A,S}^{\mathcal{MI}}$) and define the quantity of interest

$$S_{\mathrm{IAS}} := \bigcup_{S: H_{0,S}^{\mathcal{MI}} \text{ true}} S \tag{2}$$

with the convention that a union over the empty set is the empty set.

The following proposition states that S_{IAS} is a subset of the ancestors of the response Y. Similarly to PA_Y, variables in AN_Y are causes of Y in that for each ancestor there is a directed causal path to Y. Thus, generically, when intervened, these variables have a causal effect on the response.

Proposition 3.3. *It holds that $S_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$.*

The proof follows directly from (Tian et al., 1998, Theorem 2); see also (Acid & De Campos, 2013, Proposition 2). The setup in these papers is more general than what we consider here; we therefore provide direct proofs for Propositions 3.2 and 3.3 in Appendix A, which may provide further intuition for the results.

Finally, we show that the oracle output of IAS contains that of ICP and, contrary to ICP, it is always an invariant set.

Proposition 3.4. *Assume that $E \notin \mathrm{PA}_Y$. It holds that*

(i) $S_{\mathrm{IAS}} \in \mathcal{I}$ and
(ii) $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}$, with equality if and only if $S_{\mathrm{ICP}} \in \mathcal{I}$.

4. Oracle Algorithms

When provided with an oracle that tells us whether a set is invariant or not, how can we efficiently compute S_{ICP} and S_{IAS}? Here, we assume that the oracle is given by a DAG, see Assumption 2.1. A direct application of Equations (1) and (2) would require checking a number of sets that grows exponentially in the number of nodes. For S_{ICP}, we have the following characterization.[3]

Proposition 4.1. *If $E \notin \mathrm{PA}_Y$, then $S_{\mathrm{ICP}} = \mathrm{PA}_Y \cap (\mathrm{CH}_E \cup \mathrm{PA}(\mathrm{AN}_Y \cap \mathrm{CH}_E))$.*

This allows us to efficiently read off S_{ICP} from the DAG, (e.g., it can naively be done in $\mathcal{O}((d+2)^{2.373}\log(d+2))$ time, where the exponent 2.373 comes from matrix multiplication). For S_{IAS}, to the best of our knowledge, there is no closed form expression that has a similarly simple structure.

Instead, for IAS, we exploit the recent development of efficient algorithms for computing all minimal d-separators (for two given sets of nodes) in a given DAG (see, e.g., Tian et al., 1998; van der Zander et al., 2019). A set S is called a *minimal d-separator* of E and Y if it d-separates E and Y given S and no strict subset of S satisfies this property. These algorithms are often motivated by determining minimal adjustment sets (e.g., Pearl, 2009) that can be used to compute the total causal effect between two nodes, for example. If the underlying distribution is Markov and faithful with respect to the DAG, then a set S is minimally invariant if and only if it is a minimal d-separator for E and Y. We

[2] In fact, $S_{\mathrm{ICP}}^{\mathrm{MB}}$ is always at least as informative as ICP. E.g., there exist graphs in which $S_{\mathrm{ICP}} = \emptyset$ and $S_{\mathrm{ICP}}^{\mathrm{MB}} \neq \emptyset$, see Figure 1 (left). There are no possible structures for which $S_{\mathrm{ICP}}^{\mathrm{MB}} \subsetneq S_{\mathrm{ICP}}$, as both search for invariant sets over all sets of parents of Y.

[3] To the best of our knowledge, this characterization is novel.

can therefore use the same algorithms to find minimally invariant sets; van der Zander et al. (2019) provide an algorithm (based on work by Takata (2010)) for finding minimal d-separators with polynomial delay time. Applied to our case, this means that while there may be exponentially many minimally invariant sets,[4] when listing all such sets it takes at most polynomial time until the next set or the message that there are nor further sets is output. In practice, on random graphs, we found this to work well (see Section 6.1). But since S_{IAS} is the union of all minimally invariant sets, even faster algorithms may be available; to the best of our knowledge, it is an open question whether finding S_{IAS} is an NP-hard problem (see Appendix B for details).

We provide a function for listing all minimally invariant sets in our python code; it uses an implementation of the above mentioned algorithm, provided in the R (R Core Team, 2021) package `dagitty` (Textor et al., 2016). In Section 6.1, we study the properties of the oracle set S_{IAS}. When applied to 500 randomly sampled, dense graphs with $d = 15$ predictor nodes and five interventions, the `dagitty` implementation had a median speedup of a factor of roughly 17, compared to a brute-force search (over the ancestors of Y). The highest speedup achieved was by a factor of more than 1,900.

The above mentioned literature can be used only for oracle algorithms, where the graph is given. In the following sections, we discuss how to test the hypothesis of minimal invariance from data.

5. Invariant Ancestry Search

5.1. Testing a Single Set for Minimal Invariance

Usually, we neither observe a full SCM nor its graphical structure. Instead, we observe data from an SCM, which we want to use to decide whether a set is in $\mathcal{MI}$, such that we make the correct decision with high probability. We now show that a set S can be tested for minimal invariance with asymptotic level and power if given a test for invariance that has asymptotic level and power.

Assume that $\mathcal{D}_n = (X_i, E_i, Y_i)_{i=1}^n$ are observations (which may or may not be independent) of (X, E, Y) and let $\phi_n^{\mathcal{MI}} : \mathrm{powerset}([d]) \times \mathcal{D}_n \times (0,1) \to \{0,1\}$ be a decision rule that transforms $(S, \mathcal{D}_n, \alpha)$ into a decision $\phi_n^{\mathcal{MI}}(S, \mathcal{D}_n, \alpha)$ about whether the hypothesis $H_{0,S}^{\mathcal{MI}}$ should be rejected ($\phi_n^{\mathcal{MI}} = 1$) at significance threshold α, or not ($\phi_n^{\mathcal{MI}} = 0$). To ease notation, we suppress the dependence on $\mathcal{D}_n$ and α when the statements are unambiguous.

A test ψ_n for the hypothesis H_0 has pointwise asymptotic level if

$$\forall \alpha \in (0,1): \quad \sup_{\mathbb{P} \in H_0} \lim_{n \to \infty} \mathbb{P}(\psi_n = 1) \leq \alpha \tag{3}$$

and pointwise asymptotic power if

$$\forall \alpha \in (0,1): \quad \inf_{\mathbb{P} \in H_A} \lim_{n \to \infty} \mathbb{P}(\psi_n = 1) = 1. \tag{4}$$

If the limit and the supremum (resp. infimum) in Equation (3) (resp. Equation (4)) can be interchanged, we say that ψ_n has uniform asymptotic level (resp. power).

Tests for invariance have been examined in the literature. Peters et al. (2016) propose two simple methods for testing for invariance in linear Gaussian SCMs when the environments are discrete, although the methods proposed extend directly to other regression scenarios. Pfister et al. (2019) propose resampling-based tests for sequential data from linear Gaussian SCMs. Furthermore, any valid test for conditional independence between Y and E given a set of predictors S can be used to test for invariance. Although for continuous X, there exists no general conditional independence test that has both level and non-trivial power (Shah & Peters, 2020), it is possible to impose restrictions on the data-generating process that ensure the existence of non-trivial tests (e.g., Fukumizu et al., 2008; Zhang et al., 2011; Berrett et al., 2020; Shah & Peters, 2020; Thams et al., 2021). Heinze-Deml et al. (2018) provide an overview and a comparison of several conditional independence tests in the context of invariance.

To test whether a set $S \subseteq [d]$ is minimally invariant, we define the decision rule

$$\phi_n^{\mathcal{MI}}(S) := \begin{cases} 1 & \text{if } \phi_n(S) = 1 \text{ or } \min_{j \in S} \phi_n(S \setminus \{j\}) = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\phi_n^{\mathcal{MI}}(\emptyset) := \phi_n(\emptyset)$. Here, ϕ_n is a test for the hypothesis $H_{0,S}^{\mathcal{I}}$, e.g., one of the tests mentioned above. This decision rule rejects $H_{0,S}^{\mathcal{MI}}$ either if $H_{0,S}^{\mathcal{I}}$ is rejected by ϕ_n or if there exists $j \in S$ such that $H_{0,S\setminus\{j\}}^{\mathcal{I}}$ is not rejected. If ϕ_n has pointwise (resp. uniform) asymptotic level and power, then $\phi_n^{\mathcal{MI}}$ has pointwise (resp. uniform) asymptotic level and pointwise (resp. uniform) asymptotic power of at least $1 - \alpha$.

Theorem 5.1. *Let $\phi_n^{\mathcal{MI}}$ be defined as in Equation (5) and let $S \subseteq [d]$. Assume that the decision rule ϕ_n has pointwise asymptotic level and power for S and for all $S \setminus \{j\}, j \in S$. Then, $\phi_n^{\mathcal{MI}}$ has pointwise asymptotic level and pointwise asymptotic power of at least $1 - \alpha$, i.e.,*

$$\inf_{\mathbb{P} \in H_{A,S}^{\mathcal{MI}}} \lim_{n \to \infty} \mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1) \geq 1 - \alpha.$$

If ϕ_n has uniform asymptotic level and power, then $\phi_n^{\mathcal{MI}}$ has uniform asymptotic level and uniform asymptotic power of at least $1 - \alpha$.

[4] This is the case if there are $d/2$ (disjoint) directed paths between E and Y, with each path containing two X-nodes, for example (e.g., van der Zander et al., 2019).

Due to Proposition 3.3, a test for $H_{0,S}^{\mathcal{MI}}$ is implicitly a test for $S \subseteq \mathrm{AN}_Y$, and can thus be used to infer whether intervening on S will have a potential causal effect on Y. However, rejecting $H_{0,S}^{\mathcal{MI}}$ is not evidence for $S \not\subseteq \mathrm{AN}$; it is evidence for $S \notin \mathcal{MI}$.

5.2. Learning S_{IAS} from Data

We now consider the task of estimating the set S_{IAS} from data. If we are given a test for invariance that has asymptotic level and power and if we correct for multiple testing appropriately, we can estimate S_{IAS} by $\hat{S}_{\mathrm{IAS}}$, which, asymptotically, is a subset of AN_Y with large probability.

Theorem 5.2. *Assume that the decision rule ϕ_n has pointwise asymptotic level for all minimally invariant sets and pointwise asymptotic power for all $S \subseteq [d]$ such that S is not a superset of a minimally invariant set. Define $C := 2^d$ and let $\widehat{\mathcal{I}} := \{S \subseteq [d] \mid \phi_n(S, \alpha C^{-1}) = 0)\}$ be the set of all sets for which the hypothesis of invariance is not rejected and define $\widehat{\mathcal{MI}} := \left\{S \in \widehat{\mathcal{I}} \mid \forall S' \subsetneq S : S' \notin \widehat{\mathcal{I}}\right\}$ and $\hat{S}_{\mathrm{IAS}} := \bigcup_{S \in \widehat{\mathcal{MI}}} S$. It then holds that*

$$\lim_{n\to\infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y) \geq \lim_{n\to\infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}} = S_{\mathrm{IAS}}) \geq 1 - \alpha.$$

A generic algorithm for implementing $\hat{S}_{\mathrm{IAS}}$ is given in Appendix D.

Remark 5.3. Consider a decision rule ϕ_n that just (correctly) rejects the empty set (e.g., because the p-value is just below the threshold α), indicating that the effect of the environments is weak. It is likely that there are other sets $S' \notin \mathcal{I}$, which the test may not have sufficient power against and are (falsely) accepted as invariant. If one of such sets contains non-ancestors of Y, this yields a violation of $\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$. To guard against this, testing $S = \emptyset$ can be done at a lower significance level, $\alpha_0 < \alpha$. This modified IAS approach is conservative and may return $\hat{S}_{\mathrm{IAS}} = \emptyset$ if the environments do not have a strong impact on Y, but it retains the guarantee $\lim_{n\to\infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y) \geq 1 - \alpha$ of Theorem 5.2.

The multiple testing correction performed in Theorem 5.2 is strictly conservative because we only need to correct for the number of minimally invariant sets, and there does not exist 2^d minimally invariant sets. Indeed, the statement of Theorem 5.2 remains valid for $C = C'$ if the underlying DAG has at most C' minimally invariant sets. We hypothesize that a DAG can contain at most $3^{\lceil d/3 \rceil}$ minimally invariant sets and therefore propose using $C = 3^{\lceil d/3 \rceil}$ in practice. If this hypothesis is true, Theorem 5.2 remains valid (for any DAG), using $C = 3^{\lceil d/3 \rceil}$ (see Appendix C for a more detailed discussion).

Alternatively, as shown in the following section, we can restrict the search for minimally invariant sets to a predetermined size. This requires milder correction factors and comes with computational benefits.

5.3. Invariant Ancestry Search in Large Systems

We now develop a variation of Theorem 5.2, which allows us to search for ancestors of Y in large graphs, at the cost of only identifying minimally invariant sets up to some a priori determined size.

Similarly to ICP (see Section 2.2), one could restrict IAS to the variables in MB_Y but the output may be smaller than S_{IAS}; in particular, there are only non-parental ancestors in MB_Y if these are parents to both a parent a child of Y (For instance, in the graph $E \to X_1 \to \ldots \to X_d \to Y$, $S_{\mathrm{IAS}} = \{1, \ldots, d\}$ but restricting IAS to MB_Y would yield the set $\{d\}$.) Thus, we do not expect such an approach to be particularly fruitful in learning ancestors.

Here, we propose an alternative approach and define

$$S_{\mathrm{IAS}}^m := \bigcup_{S : S \in \mathcal{MI} \text{ and } |S| \leq m} S \tag{6}$$

as the union of minimally invariant sets that are no larger than $m \leq d$. For computing S_{IAS}^m, one only needs to check invariance of the $\sum_{i=0}^m \binom{d}{i}$ sets that are no larger than m. S_{IAS}^m itself, however, can be larger than m: in the graph above Equation (6), $S_{\mathrm{IAS}}^1 = \{1, \ldots, d\}$. The following proposition characterizes properties of S_{IAS}^m.

Proposition 5.4. *Let $m < d$ and let $m_{\min}$ and $m_{\max}$ be the size of a smallest and a largest minimally invariant set, respectively. The following statements are true:*

(i) $S_{\mathrm{IAS}}^m \subseteq \mathrm{AN}_Y$.
(ii) If $m \geq m_{\max}$, *then* $S_{\mathrm{IAS}}^m = S_{\mathrm{IAS}}$.
(iii) If $m \geq m_{\min}$ *and* $E \notin \mathrm{PA}_Y$, *then* $S_{\mathrm{IAS}}^m \in \mathcal{I}$.
(iv) If $m \geq m_{\min}$ *and* $E \notin \mathrm{PA}_Y$, *then* $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}^m$ *with equality if and only if* $S_{\mathrm{ICP}} \in \mathcal{I}$.

If $m < m_{\min}$ and $S_{\mathrm{ICP}} \neq \emptyset$, then $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}^m$ does not hold. However, we show in Section 6.1 using simulations that S_{IAS}^m is larger than S_{ICP} in many sparse graphs, even for $m = 1$, when few nodes are intervened on.

In addition to the computational speedup offered by considering S_{IAS}^m instead of S_{IAS}, the set S_{IAS} can be estimated from data using a smaller correction factor than the one employed in Theorem 5.2. This has the benefit that in practice, smaller sample sizes may be needed to detect non-invariance.

Theorem 5.5. *Let $m \leq d$ and define $C(m) := \sum_{i=0}^m \binom{d}{i}$. Assume that the decision rule ϕ_n has pointwise asymptotic level for all minimally invariant sets of size at most m and pointwise power for all sets of size at most m that are not supersets of a minimally invariant set. Let $\widehat{\mathcal{I}}^m := \{S \subseteq [d] \mid \phi_n(S, \alpha C(m)^{-1}) = 0$ and $|S| \leq m\}$,*

be the set of all sets of size at most m for which the hypothesis of invariance is not rejected and define $\widehat{\mathcal{MI}}^m := \left\{ S \in \widehat{\mathcal{I}}^m \mid \forall S' \subsetneq S : S' \notin \widehat{\mathcal{I}}^m \right\}$ *and* $\hat{S}^m_{\mathrm{IAS}} := \bigcup_{S \in \widehat{\mathcal{MI}}^m} S$. *It then holds that*

$$\lim_{n\to\infty} \mathbb{P}(\hat{S}^m_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y) \geq \lim_{n\to\infty} \mathbb{P}(\hat{S}^m_{\mathrm{IAS}} = S^m_{\mathrm{IAS}}) \geq 1 - \alpha.$$

The method proposed in Theorem 5.5 outputs a non-empty set if there exists a non-empty set of size at most m, for which the hypothesis of invariance cannot be rejected. In a sparse graph, it is likely that many small sets are minimally invariant, whereas if the graph is dense, it may be that all invariant sets are larger than m, such that $S^m_{\mathrm{IAS}} = \emptyset$. In dense graphs however, many other approaches may fail too; for example, it is also likely that the size of the Markov boundary is so large that applying ICP on MB_Y is not feasible.

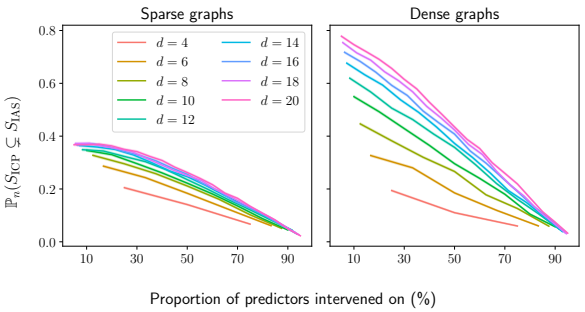


Figure 2. Low-dimensional oracle experiment, see Section 6.1. In all cases, as predicted by theory, S_{ICP} is contained in S_{IAS}. For many graphs, S_{IAS} is strictly larger than S_{ICP}. On average, this effect is more expressed when there are fewer intervened nodes. $\mathbb{P}_n$ refers to the distribution used to sample graphs and every point in the figure is based on 50,000 independently sampled graphs; d denotes the number of covariates X. Empirical confidence bands are plotted around each line, but are very narrow.

6. Experiments

We apply the methods developed in this paper in a population-case experiment using oracle knowledge (Section 6.1), a synthetic experiment using finite sample tests (Section 6.2), and a real-world data set from a gene perturbation experiment (Section 6.3). In Sections 6.1 and 6.2 we consider a setting with two environments: an observational environment ($E = 0$) and an intervention environment ($E = 1$), and examine how the strength and number of interventions affect the performance of IAS.

6.1. Oracle IAS in Random Graphs

For the oracle setting, we know that $S_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$ (Proposition 3.3) and $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}$ (Proposition 3.4). We first verify that the inclusion $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}$ is often strict in low-dimensional settings when there are few interventions. Second, we show that the set S^m_{IAS} is often strictly larger than the set $S^{\mathrm{MB}}_{\mathrm{ICP}}$ in large, sparse graphs with few interventions.

In principle, for a given number of covariates, one can enumerate all DAGs and, for each DAG, compare S_{ICP} and S_{IAS}. However, because the space of DAGs grows super-exponentially in the number of nodes (Chickering, 2002), this is infeasible. Instead, we sample graphs from the space of all DAGs that satisfy Assumption 2.1 and $Y \in \mathrm{DE}_E$ (see Appendix E.1 for details).

In the low-dimensional setting ($d \leq 20$), we compute S_{ICP} and S_{IAS}, whereas in the larger graphs ($d \geq 100$), we compute $S^{\mathrm{MB}}_{\mathrm{ICP}}$ and the reduced set S^m_{IAS} for $m \in \{1, 2\}$ when $d = 100$ and for $m = 1$ when $d = 1{,}000$. Because there is no guarantee that IAS outputs a superset of ICP when searching only up to sets of some size lower than d, we compare the size of the sets output by either method. For the low-dimensional setting, we consider both sparse and dense graphs, but for larger dimensions, we only consider sparse graphs. In the sparse setting, the DAGs are constructed such that there is an expected number of $d + 1$ edges between the $d + 1$ nodes X and Y; in the dense setting, the expected number of edges equals $0.75 \cdot d(d + 1)/2$.

The results of the simulations are displayed in Figures 2 and 3. In the low-dimensional setting, S_{IAS} is a strict superset of S_{ICP} for many graphs. This effect is the more pronounced, the larger the d and the fewer nodes are intervened on, see Figure 2. In fact, when there are interventions on all predictors, we know that $S_{\mathrm{IAS}} = S_{\mathrm{ICP}} = \mathrm{PA}_Y$ (Peters et al., 2016, Theorem 2), and thus the probability that $S_{\mathrm{ICP}} \subsetneq S_{\mathrm{IAS}}$ is exactly zero. For the larger graphs, we find that the set S^m_{IAS} is, on average, larger than $S^{\mathrm{MB}}_{\mathrm{ICP}}$, in particular when $d = 1{,}000$ or when $m = 2$, see Figure 3. In the setting with $d = 100$ and $m = 1$, the two sets are roughly the same size, when 10% of the predictors are intervened on. The set $S^{\mathrm{MB}}_{\mathrm{ICP}}$ becomes larger than S^1_{IAS} after roughly 15% of the predictors nodes are intervened on (not shown). For both $d = 100$ and $d = 1{,}000$, the average size of the Markov boundary of Y was found to be approximately 3.5.

6.2. Simulated Linear Gaussian SCMs

In this experiment, we show through simulation that IAS finds more ancestors than ICP in a finite sample setting when applied to linear Gaussian SCMs. To compare the outputs of IAS and ICP, we use the *Jaccard similarity* between $\hat{S}_{\mathrm{IAS}}$ ($\hat{S}^1_{\mathrm{IAS}}$ when d is large) and AN_Y, and between $\hat{S}_{\mathrm{ICP}}$ ($\hat{S}^{\mathrm{MB}}_{\mathrm{ICP}}$

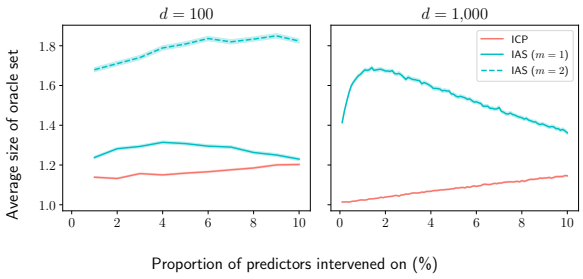


Figure 3. High-dimensional oracle experiment with sparse graphs, see Section 6.1. The average size of the set S^m_{IAS} is larger than the average size of the set $S^{\text{MB}}_{\text{ICP}}$, both when using IAS to search for sets up to sizes $m = 1$ and $m = 2$. Except for the choice of d, the setup is the same as in Figure 2.

when d is large[5]) and AN_Y.[6]

We sample data from sparse linear Gaussian models with i.i.d. noise terms in two scenarios, $d = 6$ and $d = 100$. In both cases, coefficients for the linear assignments are drawn randomly. We consider two environments; one observational and one interventional; in the interventional environment, we apply do-interventions of strength one to children of E, i.e., we fix the value of a child of E to be one. We standardize the data along the causal order, to prevent variance accumulation along the causal order (Reisach et al., 2021). Throughout the section, we consider a significance level of $\alpha = 5\%$. For a detailed description of the simulations, see Appendix E.2.

To test for invariance, we employ the test used in Peters et al. (2016): We calculate a p-value for the hypothesis of invariance of S by first linearly regressing Y onto X_S (ignoring E), and second testing whether the mean and variance of the prediction residuals is equal across environments. For details, see Peters et al. (2016, Section 3.2.1). Schultheiss et al. (2021) also consider the task of estimating ancestors but since their method is uninformative for Gaussian data and does not consider environments, it is not directly applicable here.

In Theorem 5.2, we assume asymptotic power of our invariance test. When $d = 6$, we test hypotheses with a correction factor $C = 3^{\lceil 6/3 \rceil} = 9$, as suggested in Appendix C, in an attempt to reduce false positive findings. In Appendix E.3, we repeat the experiment of this section with $C = 2^6$ and find almost identical results. We hypothesize, that the effects of a reduced C is more pronounced at larger d. When $d = 100$, we test hypotheses with the correction factor $C(1)$ of Theorem 5.5. In both cases, we test the hypothesis of invariance of the empty set at level $\alpha_0 = 10^{-6}$ (cf. Remark 5.3). In Appendix E.4, we investigate the effects on the quantities $\mathbb{P}(\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y)$ and $\mathbb{P}(\hat{S}^1_{\text{IAS}} \subseteq \text{AN}_Y)$ when varying α_0, confirming that choosing α_0 too high can lead to a reduced probability of $\hat{S}_{\text{IAS}}$ being a subset of ancestors.

In Figure 4 the results of the simulations are displayed. In SCMs where the oracle versions S_{IAS} and S_{ICP} are not equal, $\hat{S}_{\text{IAS}}$ achieved, on average, a higher Jaccard similarity to AN_Y than $\hat{S}_{\text{ICP}}$. This effect is less pronounced when $d = 100$. We believe that the difference in Jaccard similarities is more pronounced when using larger values of m. When $S_{\text{IAS}} = S_{\text{ICP}}$, the two procedures achieve roughly the same Jaccard similarities to AN_Y, as expected. When the number of observations is one hundred, IAS generally fails to find any ancestors and outputs the empty set (see Figure 7), indicating that the we do not have power to reject the empty set when there are few observations. This is partly by design; we test the empty set for invariance at reduced level α_0 in order to protect against making false positive findings when the environment has a weak effect on Y. However, even without testing the empty set at a reduced level, IAS has to correct for making multiple comparisons, contrary to ICP, thus lowering the marginal significance level each set is tested at. When computing the jaccard similarities with either $\alpha_0 = \alpha$ or $\alpha_0 = 10^{-12}$, the results were similar (not shown). We repeated the experiments with $d = 6$ with a weaker influence of the environment (do-interventions of strength 0.5 instead of 1) and found comparable results, with slightly less power in that the empty set is found more often, see Appendix E.5.

We compare our method with a variant, called $\text{IAS}_{\text{est. graph}}$, where we first estimate (e.g., using methods proposed by Mooij et al. 2020 or Squires et al. 2020) a member graph of the Markov equivalence class ('I-MEC') and apply the oracle algorithm from Section 4 (by reading of d-separations in that graph) to estimate $\mathcal{MI}$. In general, however, such an approach comes with additional assumptions; furthermore, even in the linear setup considered here, its empirical performance for large graphs is worse than the proposed method IAS, see Appendix E.7.

6.3. IAS in High Dimensional Genetic Data

We evaluate our approach in a data set on gene expression in yeast (Kemmeren et al., 2014). The data contain full-genome mRNA expressions of $d = 6{,}170$ genes and consists of $n_{\text{obs}} = 160$ unperturbed observations ($E = 0$) and $n_{\text{int}} = 1{,}479$ intervened-upon observations ($E = 1$); each of the latter observations correspond to the deletion of a single (known) gene. For each response gene $\texttt{gene}_Y \in [d]$, we apply the procedure from Section 5.3 with $m = 1$ to search

[5] $\hat{\text{MB}}$ is a Lasso regression estimate of MB_Y containing at most 10 variables

[6] The Jaccard similarity between two sets A and B is defined as $J(A, B) := |A \cap B|/|A \cup B|$, with $J(\emptyset, \emptyset) = 0$. The Jaccard similarity equals one if the two sets are equal, zero if they are disjoint and takes a value in $(0, 1)$ otherwise.

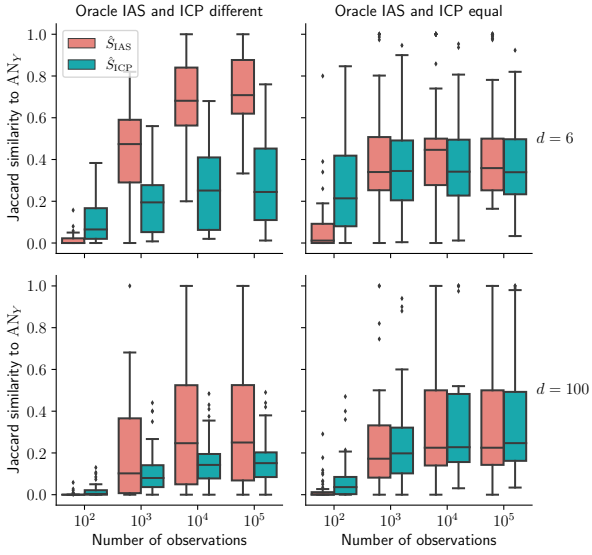


Figure 4. Comparison between the finite sample output of IAS and ICP and AN_Y on simulated data, see Section 6.2. The plots show the Jaccard similarities between AN_Y and either $\hat{S}_{\mathrm{IAS}}$ ($\hat{S}^1_{\mathrm{IAS}}$ when $d = 100$) in red or $\hat{S}_{\mathrm{ICP}}$ ($\hat{S}^{\mathrm{MB}}_{\mathrm{ICP}}$ when $d = 100$) in blue and AN_Y. When $S_{\mathrm{ICP}} \neq S_{\mathrm{IAS}}$ (left column), $\hat{S}_{\mathrm{IAS}}$ is more similar to AN_Y than $\hat{S}_{\mathrm{ICP}}$. The procedures are roughly equally similar to AN_Y when $S_{\mathrm{ICP}} = S_{\mathrm{IAS}}$ (right column). Graphs represented in each boxplot: 42 (top left), 58 (top right), 40 (bottom left) and 60 (bottom right).

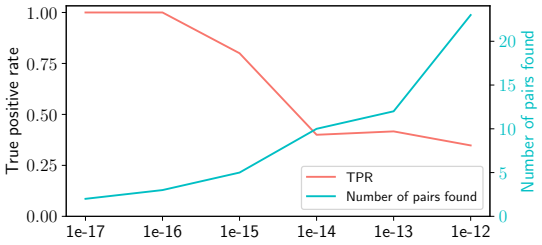


Figure 5. True positive rates and number of gene pairs found in the experiment in Section 6.3. On the x-axis, we change α_0, the threshold for invariance of the empty set. When α_0 is small, we only search for pairs if the environment has a very significant effect on Y. For smaller α_0, fewer pairs are found to be invariant (blue line), but those found, are more likely to be true positives (red line). This supports the claim that the lower α_0 is, the more conservative our approach is.

for ancestors.

We first test for invariance of the empty set, i.e., whether the distribution of $\texttt{gene}_Y$ differs between the observational and interventional environment. We test this at a conservative level $\alpha_0 = 10^{-12}$ in order to protect against a high false positive rate (see Remark 5.3). For 3,631 out of 6,170 response genes, the empty set is invariant, and we disregard them as response genes.

For each response gene, for which the empty set is not invariant, we apply our procedure. More specifically, when testing whether $\texttt{gene}_X$ is an ancestor of $\texttt{gene}_Y$, we exclude any observation in which either $\texttt{gene}_X$ or $\texttt{gene}_Y$ was intervened on. We then test whether the empty set is still rejected, at level $\alpha_0 = 10^{-12}$, and whether $\texttt{gene}_X$ is invariant at level $\alpha = 0.25$. Since a set $\{\texttt{gene}_X\}$ is deemed minimally invariant if the p-value exceeds α, setting α large is conservative for the task of finding ancestors. Indeed, when estimating $\hat{S}^m_{\mathrm{IAS}}$, one can test the sets of size m at a higher level $\alpha_1 > \alpha$. This is conservative, because falsely rejecting a minimally invariant set of size m does not break the inclusion $\hat{S}^m_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$. However, if one has little power against the non-invariant sets of size m, testing at level α_1 can protect against false positives.[7]

We use the held-out data point, where $\texttt{gene}_X$ is intervened on, to determine as ground truth, whether $\texttt{gene}_X$ is indeed an ancestor of $\texttt{gene}_Y$. We define $\texttt{gene}_X$ as a true ancestor of $\texttt{gene}_Y$ if the value of $\texttt{gene}_Y$ when $\texttt{gene}_X$ is intervened on, lies in the $q_{TP} = 1\%$ tails of the observational distribution of $\texttt{gene}_Y$.

We find 23 invariant pairs $(\texttt{gene}_X, \texttt{gene}_Y)$; of these, 7 are true positives. In comparison, Peters et al. (2016) applies ICP to the same data, and with the same definition of true positives. They predict 8 pairs, of which 6 are true positives. This difference is in coherence with the motivation put forward in Section 5.2: Our approach predicts many more ancestral pairs (8 for ICP compared to 23 for IAS). Since ICP does not depend on power of the test, they have a lower false positive rate (25% for ICP compared to 69.6% for IAS).

In Figure 5, we explore how changing α_0 and q_{TP} impacts the true positive rate. Reducing α_0 increases the true positive rate, but lowers the number of gene pairs found (see Figure 5). This is because a lower α_0 makes it more difficult to detect non-invariance of the empty set, making the procedure more conservative (with respect to finding ancestors); see Remark 5.3. For example, when $\alpha_0 \leq 10^{-15}$, the true positive rate is above 0.8; however, 5 or fewer pairs are found. When searching for ancestors, the effect of intervening may be reduced by noise from intermediary variables, so $q_{TB} = 1\%$ might be too strict; in Appendix E.6, we analyze the impact of increasing q_{TB}.

[7] Only sets of size exactly m can be tested at level α_1; the remaining hypotheses should still be corrected by $C(m)$ (or by the hypothesized number of minimally invariant sets).

7. Extensions

7.1. Latent variables

In Assumption 2.1, we assume that all variables X are observed and that there are no hidden variables H. Let us write $X = X_O \dot\cup X_H$, where only X_O is observed and define $\mathcal{I} \coloneqq \{S \subseteq X_O \mid S \text{ invariant}\}$. We can then define

$$S_{\mathrm{IAS},O} \coloneqq \bigcup_{S \subseteq X_O : H_{0,S}^{\mathcal{M}_\mathcal{I}} \text{ true}} S$$

(again with the convention that a union over the empty set is the empty set), and have the following modification of Proposition 3.3.

Proposition 7.1. *It holds that* $S_{\mathrm{IAS},O} \subseteq \mathrm{AN}_Y$.

All results in this paper remain correct in the presence of hidden variables, except for Proposition 3.4 and Proposition 5.4 (iii-iv).[8] Thus, the union of the observed minimally invariant sets, $S_{\mathrm{IAS},O}$ is a subset of AN_Y and can be learned from data in the same way as if no latent variables were present.

7.2. Non-exogenous environments

Throughout this paper, we have assumed that the environment variable is exogenous (Assumption 2.1). However, all of the results stated in this paper, except for Proposition 4.1, also hold under the alternative assumption that E is an ancestor of Y, but not necessarily exogenous. From the remaining results, only the proof of Proposition 3.2 uses exogeneity of E, but here the result follows from Tian et al. (1998). In all other proofs, we account for both options. This extension also remains valid in the presence of hidden variables, using the same arguments as in Section 7.1.

8. Conclusion and Future Work

Invariant Ancestry Search (IAS) provides a framework for searching for causal ancestors of a response variable Y through finding minimally invariant sets of predictors by exploiting the existence of exogenous heterogeneity. The set S_{IAS} is a subset of the ancestors of Y, a superset of S_{ICP} and, contrary to S_{ICP}, invariant itself. Furthermore, the hierarchical structure of minimally invariant sets allows IAS to search for causal ancestors only among subsets up to a predetermined size. This avoids exponential runtime and allows us to apply the algorithm to large systems. We have shown that, asymptotically, S_{IAS} can be identified from data with high probability if we are provided with a test for invariance that has asymptotic level and power. We have validated our procedure both on simulated and real data. Our proposed framework would benefit from further research in the maximal number of minimally invariant sets among graphs of a fixed size, as this would provide larger finite sample power for identifying ancestors. Further it is of interest to establish finite sample guarantees or convergence rates for IAS, possibly by imposing additional assumptions on the class of SCMs. Finally, even though current implementations are fast, it is an open theoretical question whether computing S_{IAS} in the oracle setting of Section 4 is NP-hard, see Appendix B.

[8] These results do not hold in the presence of hidden variables, because it is not guaranteed that an invariant set exists among X_O (e.g., consider a graph where all observed variables share a common, unobserved confounder with Y). However, if at least one minimally invariant set exists among the observed variables, then all results stated in this paper hold.

Acknowledgements

NT and JP were supported by a research grant (18968) from VILLUM FONDEN.

References

Acid, S. and De Campos, L. M. An algorithm for finding minimum d-separating sets in belief networks. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1): 175–197, 2020.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. Foundations of structural causal models with cycles and latent variables. *Annals of Statistics*, 49(5):2885–2915, 2021.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3: 507–554, 2002.

Christiansen, R., Pfister, N., Jakobsen, M. E., Gnecco, N., and Peters, J. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (accepted)*, 2022.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20, 2008.

Gamella, J. L. and Heinze-Deml, C. Active invariant causal prediction: Experiment selection through stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Gaspers, S. and Mackenzie, S. On the number of minimal separators in graphs. In *International Workshop on Graph-Theoretic Concepts in Computer Science*, pp. 116–121. Springer, 2015.

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 21, 2009.

Jordan, M. I. Artificial intelligence — the revolution hasn't happened yet. *Harvard Data Science Review*, 1(1), 2019.

Kemmeren, P., Sameith, K., Van De Pasch, L. A., Benschop, J. J., Lenstra, T. L., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C. W., et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157 (3):740–752, 2014.

Lauritzen, S. L. *Graphical models*. Clarendon Press, 1996.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. Domain adaptation by using causal inference to predict invariant conditional distributions. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 10846–10856. Curran Associates, Inc., 2018.

Martinet, G., Strzalkowski, A., and Engelhardt, B. E. Variance minimization in the Wasserstein space for invariant causal prediction. *arXiv preprint arXiv:2110.07064*, 2021.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

Mooij, J. M., Magliacane, S., and Claassen, T. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020. URL http://jmlr.org/papers/v21/17-123.html.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. The Morgan Kaufmann series in representation and learning. Morgan Kaufmann, 2014.

Pearl, J. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

Pfister, N., Bühlmann, P., and Peters, J. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Reisach, A., Seiler, C., and Weichwald, S. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Schultheiss, C., Bühlmann, P., and Yuan, M. Higher-order least squares: assessing partial goodness of fit of linear regression. *arXiv preprint arXiv:2109.14544*, 2021.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3):1514–1538, 2020.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

Squires, C., Wang, Y., and Uhler, C. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR, 2020.

Takata, K. Space-optimal, backtracking algorithms to list the minimal vertex separators of a graph. *Discrete Applied Mathematics*, 158:1660–1667, 2010.

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., and Ellison, G. T. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, 2016.

Thams, N., Saengkyongam, S., Pfister, N., and Peters, J. Statistical testing under distributional shifts. *arXiv preprint arXiv:2105.10821*, 2021.

Tian, J., Paz, A., and Pearl, J. Finding minimal d-separators. Technical report, University of California, Los Angeles, 1998.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

van der Zander, B., Liśkiewicz, M., and Textor, J. Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework. *Artificial Intelligence*, 270:1–40, 2019.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 804–813, 2011.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

A. Proofs

A.1. A direct Proof of Proposition 3.2

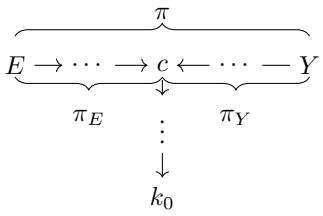
Proof. Assume that E is exogenous. If $E \in \mathrm{PA}_Y$, then there are no minimally invariant sets, and the statement holds trivially. If $E \notin \mathrm{PA}_Y$, then assume for contradiction, that an invariant set $S_0 \subsetneq S$ exists. By assumption, $|S \setminus S_0| > 1$, because otherwise S_0 would be non-invariant.

We can choose $S_1 \subseteq S$ and $k_0, k_1, \ldots, k_l \in S$ with $l \geq 1$ such that for all $i = 1, \ldots, l : k_i \notin \mathrm{DE}_{k_0}$ and

$$\begin{aligned} & S_0 \cup S_1 \cup \{k_0, \ldots, k_l\} = S && \in \mathcal{I} \\ \text{for } 0 \leq i < l: \quad & S_0 \cup S_1 \cup \{k_0, \ldots, k_i\} && \notin \mathcal{I} \\ & S_0 \cup S_1 && \in \mathcal{I}. \end{aligned}$$

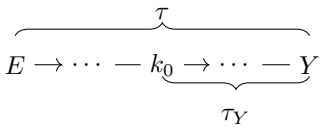
This can be done by iteratively removing elements from $S \setminus S_0$, removing first the earliest elements in the causal order. The first invariant set reached in this process is then $S_0 \cup S_1$.

Since $S_0 \cup S_1 \cup \{k_0\}$ is non-invariant, there exists a path π between E and Y that is open given $S_0 \cup S_1 \cup \{k_0\}$ but blocked given $S_0 \cup S_1$. Since removing k_0 blocks π, k_0 must be a collider or a descendant of a collider c on π:

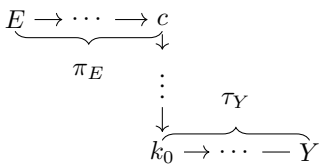


Here, $-$ represents an edge that either points left or right. Since π is open given $S_0 \cup S_1$, the two sub-paths π_E and π_Y are open given $S_0 \cup S_1$.

Additionally, since $S_0 \cup S_1 \cup \{k_1, \ldots, k_l\} = S \setminus \{k_0\}$ is non-invariant, there exists a path τ between E and Y that is unblocked given $S_0 \cup S_1 \cup \{k_1, \ldots, k_l\}$ and blocked given $S_0 \cup S_1 \cup \{k_1, \ldots, k_l\} \cup \{k_0\}$. It follows that k_0 lies on τ (otherwise τ cannot be blocked by adding k_0) and k_0 has at least one outgoing edge. Assume, without loss of generality that there is an outgoing edge towards Y. Since τ is open given $S_0 \cup S_1 \cup \{k_1, \ldots, k_l\}$, so is τ_Y.



If there are no colliders on τ_Y, then τ_Y is also open given $S_0 \cup S_1$. But then the path the path $E \overset{\pi_E}{\cdots\!\rightarrow} c \rightarrow \cdots \rightarrow k_0 \overset{\tau_Y}{\rightarrow\!\cdots}$ is also open given $S_0 \cup S_1$, contradicting invariance of $S_0 \cup S_1$.



If there are colliders on τ_Y, let m be the collider closest to k_0, meaning that $m \in \mathrm{DE}_{k_0}$. Since τ_Y is open given $S_0 \cup S_1 \cup \{k_1, \ldots, k_l\}$, it means that either m or a descendant of m is in $S_0 \cup S_1 \cup \{k_1, \ldots, k_l\}$. Since $\{k_1, \ldots, k_l\} \cap \mathrm{DE}_{k_0} = \emptyset$, there exist $v \in (S_0 \cup S_1) \cap (\{m\} \cup \mathrm{DE}_m)$. But then $v \in \mathrm{DE}_{k_0} \cap (S_0 \cup S_1)$, meaning that π is open given $S_0 \cup S_1$, contradicting invariance of $S_0 \cup S_1$.

We could assume that τ_Y had an outgoing edge from k_0 without loss of generality, because if there was instead an outgoing edge from k_0 on τ_E, the above argument would work with π_Y and τ_E instead. This concludes the proof. □

A.2. A direct proof of Proposition 3.3

Proof. If E is a parent of Y, we have $\mathcal{MI} = \emptyset$ and the statement follows trivially. Thus, assume that E is not a parent of Y. We will show that if $S \in \mathcal{I}$ is not a subset of AN_Y, then $S^* := S \cap \mathrm{AN}_Y \in \mathcal{I}$, meaning that $S \notin \mathcal{MI}$.

Assume for contradiction that there is a path p between E and Y that is open given S^*. Since $S \in \mathcal{I}$, p is blocked given S. Then there exists a non-collider Z on p that is in $S \setminus \mathrm{AN}_Y$. We now argue that all nodes on p are ancestors of Y, yielding a contradiction.

First, assume that there are no colliders on p. If E is exogenous, then p is directed from E to Y. (If E is an ancestor of Y, any node on p is either an ancestor of Y or E, and thus Y.) Second, assume that there are colliders on p. Since p is open given the smaller set $S^* \subsetneq S$, all colliders on p are in S^* or have a descendant in S^*; therefore all colliders are ancestors of Y. If E is exogenous, any node on p is either an ancestor of Y or of a collider on p. (If E is an ancestor of Y, any node on p is either an ancestor of Y, of a collider on p or of E, and thus also Y.) This completes the proof of Proposition 3.3.

□

A.3. Proof of Proposition 3.4

Proof. First, we show that $S_{\mathrm{IAS}} \in \mathcal{I}$. If S_{IAS} is the union of a single minimally invariant set, it trivially holds that $S_{\mathrm{IAS}} \in \mathcal{I}$. Now assume that S_{IAS} is the union of at least two minimally invariant sets, $S_{\mathrm{IAS}} = S_1 \cup \ldots \cup S_n$, $n \geq 2$, and assume for a contradiction that there exists a path π between E and Y that is unblocked given S_{IAS}.

Since π is blocked by a strict subset of S_{IAS}, it follows that π has at least one collider; further every collider of π is either in S_{IAS} or has a descendant in S_{IAS}, and hence every collider of π is an ancestor of Y, by Proposition 3.3. If E is exogenous, π has the following shape

$$E \overbrace{\longrightarrow \cdots \longrightarrow}^{\pi_1} c_1 \overbrace{\longleftarrow \cdots \longrightarrow}^{\pi_2} c_2 \overbrace{\longleftarrow \cdots \longrightarrow}^{\pi_3, \ldots, \pi_k} c_k \overbrace{\longleftarrow \cdots \longrightarrow}^{\pi_{k+1}} Y.$$

(If E is not exogenous but $E \in \mathrm{AN}_Y$, then π takes either the form displayed above or the shape displayed below. However,

$$E \overbrace{\longleftarrow \cdots \longrightarrow}^{\pi_1} c_1 \overbrace{\longleftarrow \cdots \longrightarrow}^{\pi_2} c_2 \overbrace{\longleftarrow \cdots \longrightarrow}^{\pi_3, \ldots, \pi_k} c_k \overbrace{\longleftarrow \cdots \longrightarrow}^{\pi_{k+1}} Y.$$

no matter which of the shapes π takes, the proof proceeds the same.) The paths $\pi_1, \ldots, \pi_{k+1}$, $k \geq 1$, do not have any colliders and are unblocked given S_{IAS}. In particular, $\pi_1, \ldots, \pi_{k+1}$ are unblocked given S_1.

The path π_{k+1} must have a final edge pointing to Y, because otherwise it would be a directed path from Y to c_k, which contradicts acyclicity since c_k is an ancestor of Y.

As c_1 is an ancestor of Y, there exists a directed path, say ρ_1, from c_1 to Y. Since π_1 is open given S_1 and since S_1 is invariant, it follows that ρ_1 must be blocked by S_1 (otherwise the path $E \overset{\pi_1}{\to} c_1 \overset{\rho_1}{\to} Y$ would be open). For this reason, S_1 contains a descendant of the collider c_1.

Similarly, if ρ_2 is a directed path from c_2 to Y, then S_1 blocks ρ_2, because otherwise the path $E \overset{\pi_1}{\to} c_1 \leftarrow \overset{\pi_2}{\cdots} \to c_2 \overset{\rho_2}{\to} Y$ would be open. Again, for this reason, S_1 contains a descendant of c_2.

Iterating this argument, it follows that S_1 contains a descendant of every collider on π, and since $\pi_1, \ldots, \pi_{k+1}$ are unblocked by S_1, π is open given S_1. This contradicts invariance of S_1 and proves that $S_{\mathrm{IAS}} \in \mathcal{I}$.

We now show that $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}$ with equality if and only if $S_{\mathrm{ICP}} \in \mathcal{I}$. First, $S_{\mathrm{ICP}} \subseteq S_{\mathrm{IAS}}$ because S_{IAS} is a union of the minimally invariant sets, and S_{ICP} is the intersection over all invariant sets. We now show the equivalence statement.

Assume first that $S_{\mathrm{ICP}} \in \mathcal{I}$. As S_{ICP} is the intersection of all invariant sets, $S_{\mathrm{ICP}} \in \mathcal{I}$ implies that there exists exactly one

invariant set, that is contained in all other invariant sets. By definition, this means that there is only one minimally invariant set, and that this set is exactly S_{ICP}. Thus, $S_{\text{IAS}} = S_{\text{ICP}}$.

Conversely assume that $S_{\text{ICP}} \notin \mathcal{I}$. By construction, S_{ICP} is contained in any invariant set, in particular in the minimally invariant sets. However, since S_{ICP} is not invariant itself, this containment is strict, and it follows that $S_{\text{ICP}} \subsetneq S_{\text{IAS}}$.

□

A.4. Proof of Proposition 4.1

Proof. First we show $\text{PA}_Y \cap (\text{CH}_E \cup \text{PA}(\text{AN}_Y \cap \text{CH}_E)) \subseteq S_{\text{ICP}}$. If $j \in \text{PA}_Y \cap \text{CH}_E$, any invariant set contains j, because otherwise the path $E \to j \to Y$ is open. Similarly, if $j \in \text{PA}_Y \cap \text{PA}(\text{AN}_Y \cap \text{CH}_E)$, any invariant set contains j (there exists a node j' such that $E \to j' \to \cdots \to Y$ and $E \to j' \leftarrow j \to Y$, and any invariant set S must contain j' or one of its descendants; thus, it must also contain j to ensure that the path $E \to j' \leftarrow j \to Y$ is blocked by S.) It follows that for all invariant S,

$$\text{PA}_Y \cap (\text{CH}_E \cup \text{PA}(\text{AN}_Y \cap \text{CH}_E)) \subseteq S,$$

such that

$$\text{PA}_Y \cap (\text{CH}_E \cup \text{PA}(\text{AN}_Y \cap \text{CH}_E)) \subseteq \bigcap_{S \text{ invariant}} S.$$

To show $S_{\text{ICP}} \subseteq \text{PA}_Y \cap (\text{CH}_E \cup \text{PA}(\text{AN}_Y \cap \text{CH}_E))$, take any $j \notin \text{PA}_Y \cap (\text{CH}_E \cup \text{PA}(\text{AN}_Y \cap \text{CH}_E))$. We argue, that an invariant set $\bar{S}$ not containing j exists, such that $j \notin S_{\text{ICP}} = \bigcap_{S \text{ invariant}} S$. If $j \notin \text{PA}_Y$, let $\bar{S} = \text{PA}_Y$, which is invariant. If $j \in \text{PA}_Y$, define

$$\bar{S} = (\text{PA}_Y \setminus \{j\}) \cup \text{PA}_j \cup (\text{CH}_j \cap \text{AN}_Y) \cup \text{PA}(\text{CH}_j \cap \text{AN}_Y).$$

Because $j \notin \text{CH}_E$ and $j \notin \text{PA}(\text{AN}_Y \cap \text{CH}_E)$, we have $E \notin \bar{S}$. Also observe that $\bar{S} \subseteq \text{AN}_Y$. We show that any path between E and Y is blocked by $\bar{S}$, by considering all possible paths:

$\cdots \mathbf{j'} \to \mathbf{Y}$ **for** $\mathbf{j'} \neq \mathbf{j}$**:** Blocked because $j' \in \text{PA}_Y \setminus \{j\}$.

$\cdots \mathbf{v} \to \mathbf{j} \to \mathbf{Y}$**:** Blocked because $v \in \text{PA}_j \subseteq \bar{S}$ and $E \notin \text{PA}_j$.

$\cdots \mathbf{v} \to \mathbf{c} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \in \text{AN}_{\mathbf{Y}}$**:** Blocked because $v \in \text{PA}_j(\text{CH}_j \cap \text{AN}_Y)$.

$\cdots \mathbf{v} \to \mathbf{c} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \notin \text{AN}_{\mathbf{Y}}$**:** Blocked because $\bar{S} \subseteq \text{AN}_Y$, and since $c \notin \text{AN}_Y$, $\bar{S} \cap \text{DE}_c = \emptyset$ and the path is blocked given $\bar{S}$ because of the collider c.

$\cdots \to \mathbf{c} \leftarrow \cdots \leftarrow \mathbf{v} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \in \text{AN}_{\mathbf{Y}}$**:** Blocked because $v \in \text{AN}_c$ and $c \in \text{AN}_Y$, so $v \in \text{CH}_j \cap \text{AN}_Y \subseteq \bar{S}$.

$\cdots \to \mathbf{c} \leftarrow \cdots \leftarrow \mathbf{v} \leftarrow \mathbf{j} \to \mathbf{Y}$ **and** $\mathbf{c} \notin \text{AN}_{\mathbf{Y}}$**:** Same reason as for the case '$\cdots \mathbf{v} \to \mathbf{c} \leftarrow \mathbf{j} \to \mathbf{Y}$ and $\mathbf{c} \notin \text{AN}_{\mathbf{Y}}$'.

$\cdots \to \mathbf{c} \leftarrow \cdots \leftarrow \mathbf{Y}$ Since $\bar{S} \subseteq \text{AN}_Y$, we must have $\bar{S} \cap \text{DE}_c = \emptyset$ (otherwise this would create a directed cycle from $Y \to \cdots \to Y$). Hence the path is blocked given $\bar{S}$ because of the collider c.

Since there are no open paths from E to Y given $\bar{S}$, $\bar{S}$ is invariant, and $S_{\text{ICP}} \subseteq \bar{S}$. Since $j \notin \bar{S}$, it follows that $j \notin S_{\text{ICP}}$. This concludes the proof. □

A.5. Proof of Theorem 5.1

Proof. Consider first the case where all marginal tests have pointwise asymptotic power and pointwise asymptotic level.

Pointwise asymptotic level: Let $\mathbb{P}_0 \in H_{0,S}^{\mathcal{M}^{\mathcal{I}}}$. By the assumption of pointwise asymptotic level, there exists a non-negative

sequence $(\epsilon_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty}\epsilon_n = 0$ and $\mathbb{P}_0(\phi_n(S) = 1) \leq \alpha + \epsilon_n$. Then

$$\begin{aligned}
\mathbb{P}_0(\phi_n^{\mathcal{MI}}(S) = 1) &= \mathbb{P}_0\left((\phi_n(S) = 1) \cup \bigcup_{j\in S}(\phi_n(S\setminus\{j\}) = 0)\right) \\
&\leq \mathbb{P}_0(\phi_n(S) = 1) + \sum_{j\in S}\mathbb{P}_0(\phi_n(S\setminus\{j\}) = 0) \\
&\leq \alpha + \epsilon_n + \sum_{j\in S}\mathbb{P}_0(\phi_n(S\setminus\{j\}) = 0) \\
&\to \alpha + 0 \qquad \text{as } n\to\infty \\
&= \alpha.
\end{aligned}$$

The convergence step follows from

$$H_{0,S}^{\mathcal{MI}} = H_{0,S}^{\mathcal{I}} \cap \bigcap_{j\in S} H_{A,S\setminus\{j\}}^{\mathcal{I}}$$

and from the assumption of pointwise asymptotic level and power. As $\mathbb{P}_0 \in H_{0,S}^{\mathcal{MI}}$ was arbitrary, this shows that $\phi_n^{\mathcal{MI}}$ has pointwise asymptotic level.

Pointwise asymptotic power: To show that the decision rule has pointwise asymptotic power, consider any $\mathbb{P}_A \in H_{A,S}^{\mathcal{MI}}$. We have that

$$H_{A,S}^{\mathcal{MI}} = H_{A,S}^{\mathcal{I}} \cup \left(H_{0,S}^{\mathcal{I}} \cap \bigcup_{j\in S} H_{0,S\setminus\{j\}}^{\mathcal{I}}\right). \tag{7}$$

As the two sets $H_{A,S}^{\mathcal{I}}$ and

$$H_{0,S}^{\mathcal{I}} \cap \bigcup_{j\in S} H_{0,S\setminus\{j\}}^{\mathcal{I}}$$

are disjoint, we can consider them one at a time. Consider first the case $\mathbb{P}_A \in H_{A,S}^{\mathcal{I}}$. This means that S is not invariant and thus

$$\begin{aligned}
\mathbb{P}_A(\phi_n^{\mathcal{MI}}(S) = 1) &= \mathbb{P}_A\left((\phi_n(S) = 1) \cup \bigcup_{j\in S}(\phi_n(S\setminus\{j\}, \alpha) = 0)\right) \\
&\geq \mathbb{P}_A(\phi_n(S) = 1) \\
&\to 1 \qquad \text{as } n\to\infty
\end{aligned}$$

by the assumption of pointwise asymptotic power.

Next, assume that there exists $j' \in S$ such that $\mathbb{P}_A \in (H_{0,S}^{\mathcal{I}} \cap H_{0,S\setminus\{j'\}}^{\mathcal{I}})$. Then,

$$\begin{aligned}
\mathbb{P}_A(\phi_n^{\mathcal{MI}}(S) = 1) &= \mathbb{P}_0\left((\phi_n(S) = 1) \cup \bigcup_{j\in S}(\phi_n(S\setminus\{j\}) = 0)\right) \\
&\geq \mathbb{P}_A(\phi_n(S\setminus\{j'\}) = 0) \\
&\geq 1 - \alpha - \epsilon_n \\
&\to 1 - \alpha \qquad \text{as } n\to\infty.
\end{aligned}$$

Thus, for arbitrary $\mathbb{P}_A \in H_{A,S}^{\mathcal{MI}}$ we have shown that $\mathbb{P}_A(\phi_n^{\mathcal{MI}}(S) = 1) \geq 1 - \alpha$ in the limit. This shows that $\phi_n^{\mathcal{MI}}$ has pointwise asymptotic power of at least $1 - \alpha$. This concludes the argument for pointwise asymptotic power.

Next, consider the case that the marginal tests have uniform asymptotic power and uniform asymptotic level. The calculations for showing that $\phi_n^{\mathcal{MI}}$ has uniform asymptotic level and uniform asymptotic power of at least $1 - \alpha$ are almost identical to the pointwise calculations.

Uniform asymptotic level: By the assumption of uniform asymptotic level, there exists a non-negative sequence ϵ_n such that $\lim_{n\to\infty} \epsilon_n = 0$ and $\sup_{\mathbb{P}\in H^{\mathcal{I}}_{0,S}} \mathbb{P}(\phi_n(S) = 1) \leq \alpha + \epsilon_n$. Then,

$$\begin{aligned}
\sup_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1) &= \sup_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}} \mathbb{P}\left((\phi_n(S) = 1) \cup \bigcup_{j\in S} (\phi_n(S \setminus \{j\}) = 0)\right) \\
&\leq \sup_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}} \left(\mathbb{P}(\phi_n(S) = 1) + \sum_{j\in S} \mathbb{P}(\phi_n(S \setminus \{j\}) = 0)\right) \\
&\leq \sup_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}} \mathbb{P}(\phi_n(S) = 1) + \sum_{j\in S} \sup_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}} \mathbb{P}(\phi_n(S \setminus \{j\}) = 0) \\
&\leq \alpha + \epsilon_n + \sum_{j\in S} \left(1 - \inf_{\mathbb{P}\in H^{\mathcal{MI}}_{0,S}} \mathbb{P}(\phi_n(S \setminus \{j\}) = 1)\right) \\
&\to \alpha + 0 + \sum_{j\in S} (1 - 1) \qquad \text{as } n \to \infty \\
&= \alpha.
\end{aligned}$$

Uniform asymptotic power: From (7), it follows that

$$\inf_{\mathbb{P}\in H^{\mathcal{MI}}_{A,S}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1) = \min\left\{\inf_{\mathbb{P}\in H^{\mathcal{I}}_{A,S}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1), \inf_{\mathbb{P}\in H^{\mathcal{I}}_{0,S} \cap \bigcup_{j\in S} H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1)\right\}.$$

We consider the two inner terms in the above separately. First,

$$\begin{aligned}
\inf_{\mathbb{P}\in H^{\mathcal{I}}_{A,S}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1) &= \inf_{\mathbb{P}\in H^{\mathcal{I}}_{A,S}} \mathbb{P}\left((\phi_n(S) = 1) \cup \bigcup_{j\in S} (\phi_n(S \setminus \{j\}) = 0)\right) \\
&\geq \inf_{\mathbb{P}\in H^{\mathcal{I}}_{A,S}} \mathbb{P}(\phi_n(S) = 1) \\
&\to 1 \qquad \text{as } n \to \infty.
\end{aligned}$$

Next,

$$\begin{aligned}
\inf_{\mathbb{P}\in H^{\mathcal{I}}_{0,S} \cap \bigcup_{j\in S} H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}(\phi_n^{\mathcal{MI}}(S) = 1) &= \inf_{\mathbb{P}\in H^{\mathcal{I}}_{0,S} \cap \bigcup_{j\in S} H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}\left((\phi_n(S) = 1) \cup \bigcup_{j\in S} (\phi_n(S \setminus \{j\}) = 0)\right) \\
&= \min_{j\in S}\left\{\inf_{\mathbb{P}\in H^{\mathcal{I}}_{0,S} \cap H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}\left((\phi_n(S) = 1) \cup \bigcup_{j\in S} (\phi_n(S \setminus \{j\}) = 0)\right)\right\} \\
&\geq \min_{j\in S}\left\{\inf_{\mathbb{P}\in H^{\mathcal{I}}_{0,S} \cap H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}(\phi_n(S \setminus \{j\}) = 0)\right\} \\
&= \min_{j\in S}\left\{1 - \sup_{\mathbb{P}\in H^{\mathcal{I}}_{0,S} \cap H^{\mathcal{I}}_{0,S\setminus\{j\}}} \mathbb{P}(\phi_n(S \setminus \{j\}) = 1)\right\} \\
&\geq 1 - \alpha - \epsilon_n \\
&\to 1 - \alpha \qquad \text{as } n \to \infty.
\end{aligned}$$

This shows that $\phi_n^{\mathcal{MI}}$ has uniform asymptotic power of at least $1 - \alpha$, which completes the proof. □

A.6. Proof of Theorem 5.2

Proof. We have that

$$\lim_{n\to\infty} \mathbb{P}(\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y) \geq \lim_{n\to\infty} \mathbb{P}(\hat{S}_{\text{IAS}} = S_{\text{IAS}})$$

as $S_{\text{IAS}} \subseteq \text{AN}_Y$ by Proposition 3.4. Furthermore, we have

$$\mathbb{P}(\hat{S}_{\text{IAS}} = S_{\text{IAS}}) \geq \mathbb{P}(\widehat{\mathcal{MI}} = \mathcal{MI}).$$

Let $A := \{S \mid S \notin \mathcal{I}\} \setminus \{S \mid \exists S' \subsetneq S \text{ s.t. } S' \in \mathcal{MI}\}$ be those non-invariant sets that do not contain a minimally invariant set and observe that

$$(\widehat{\mathcal{MI}} = \mathcal{MI}) \supseteq \bigcap_{S\in\mathcal{MI}} (\phi_n(S, \alpha C^{-1}) = 0) \cap \bigcap_{S\in A} (\phi_n(S, \alpha C^{-1}) = 1). \tag{8}$$

To see why this is true, note that to correctly recover $\mathcal{MI}$, we need to 1) accept the hypothesis of minimal invariance for all minimally invariant sets and 2) reject the hypothesis of invariance for all non-invariant sets that are not supersets of a minimally invariant set (any superset of a set for which the hypothesis of minimal invariance is not rejected is removed in the computation of $\widehat{\mathcal{MI}}$). Then,

$$\begin{aligned}
\mathbb{P}(\widehat{\mathcal{MI}} = \mathcal{MI}) &\geq \mathbb{P}\left(\bigcap_{S\in\mathcal{MI}} (\phi_n(S, \alpha C^{-1}) = 0) \cap \bigcap_{S\in A} (\phi_n(S, \alpha C^{-1}) = 1)\right) \\
&\geq 1 - \mathbb{P}\left(\bigcup_{S\in\mathcal{MI}} (\phi_n(S, \alpha C^{-1}) = 1)\right) - \sum_{S\in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - \sum_{S\in\mathcal{MI}} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 1) - \sum_{S\in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - \sum_{S\in\mathcal{MI}} (\alpha C^{-1} + \epsilon_{n,S}) - \sum_{S\in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - |\mathcal{MI}|\alpha C^{-1} + \sum_{S\in\mathcal{MI}} \epsilon_{n,S} - \sum_{S\in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\geq 1 - \alpha + \sum_{S\in\mathcal{MI}} \epsilon_{n,S} - \sum_{S\in A} \mathbb{P}(\phi_n(S, \alpha C^{-1}) = 0) \\
&\to 1 - \alpha \quad \text{as } n \to \infty,
\end{aligned}$$

where $(\epsilon_{n,S})_{n\in\mathbb{N},S\in\mathcal{MI}}$ are non-negative sequences that converge to zero and the last step follows from the assumption of asymptotic power. The sequences $(\epsilon_{n,S})_{n\in\mathbb{N},S\in\mathcal{MI}}$ exist by the assumption of asymptotic level. □

A.7. Proof of Proposition 5.4

Proof. We prove the statements one by one.

(i) Since S^m_{IAS} is the union over some of the minimally invariant sets, $S^m_{\text{IAS}} \subseteq S_{\text{IAS}}$. Then the statement follows from Proposition 3.3.

(ii) If $m \geq m_{\max}$, all $S \in \mathcal{MI}$ satisfy the requirement $|S| \leq m$.

(iii) If $m \geq m_{\min}$, then S^m_{IAS} contains at least one minimally invariant set. The statement then follows from the first part of the proof of Proposition 3.4 given in Appendix A.3.

(iv) S^m_{IAS} contains at least one minimally invariant set and, by (iii), it is itself invariant. Thus, if $S_{\text{ICP}} \notin \mathcal{I}$, then $S_{\text{ICP}} \subsetneq S^m_{\text{IAS}}$. If $S_{\text{ICP}} \in \mathcal{I}$, then there exists only one minimally invariant set, which is S_{ICP} (see proof of Proposition 3.4), and we have $S_{\text{ICP}} = S^m_{\text{IAS}}$. This concludes the proof. □

A.8. Proof of Theorem 5.5

Proof. The proof is identical to the proof of Theorem 5.2, when changing the correction factor 2^{-d} to $C(m)^{-1}$, adding superscript m's to the quantities $\widehat{\mathcal{MI}}$, $\hat{S}_{\mathrm{IAS}}$ and S_{IAS}, and adding the condition $|S| \leq m$ to all unions, intersections and sums. □

A.9. Proof of Proposition 7.1

By Proposition 3.3, we have $S_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y$, and since $S_{\mathrm{IAS},O} \subseteq S_{\mathrm{IAS}}$, the claim follows immediately.

B. Oracle Algorithms for Learning S_{IAS}

In this section, we review some of the existing literature on minimal d-separators, which can be exploited to give an algorithmic approach for finding S_{IAS} from a DAG. We first introduce the concept of M-minimal separation with respect to a constraining set I.

Definition B.1 (van der Zander et al. (2019), Section 2.2)**.** Let $I \subseteq [d]$, $K \subseteq [d]$, and $S \subseteq [d]$. We say that S is a *K-minimal separator* of E and Y with respect to a constraining set I if all of the following are true:

(i) $I \subseteq S$.
(ii) $S \in \mathcal{I}$.
(iii) There does not exists $S' \in \mathcal{I}$ such that $K \subseteq S' \subsetneq S$.

We denote by $M_{K,I}$ the set of all K-minimal separating sets with respect to constraining set I.

(In this work, $S \in \mathcal{I}$ means $E \perp\!\!\!\perp Y \mid S$, but it can stand for other separation statements, too.) The definition of a K-minimal separator coincides with the definition of a minimally invariant set if both K and the constraining set I are equal to the empty set. An $\emptyset$-minimal separator with respect to constraining set I is called a *strongly-minimal separator with respect to constraining set I*.

We can now represent (2) using this notation. $M_{\emptyset,\emptyset}$ contains the minimally invariant sets and thus

$$S_{\mathrm{IAS}} \coloneqq \bigcup_{S \in M_{\emptyset,\emptyset}} S.$$

Listing the set $M_{I,I}$ of all I-minimal separators with respect to the constraining set I (for any I) can be done in polynomial delay time $\mathcal{O}(d^3)$ (van der Zander et al., 2019; Takata, 2010), where delay here means that finding the next element of $M_{I,I}$ (or announcing that there is no further element) has cubic complexity. This is the algorithm we exploit, as described in the main part of the paper.

Furthermore, we have

$$i \in S_{\mathrm{IAS}} \quad \Leftrightarrow \quad M_{\emptyset,\{i\}} \neq \emptyset.$$

This is because $i \in S_{\mathrm{IAS}}$ if and only if there is a minimally invariant set that contains i, which is the case if and only if there exist a strongly minimal separating set with respect to constraining set $\{i\}$. Thus, we can construct S_{IAS} by checking, for each i, whether there is an element in $M_{\emptyset,\{i\}}$. Finding a strongly-minimal separator with respect to constraining set I, i.e., finding an element in $M_{\emptyset,I}$, is NP-hard if the set I is allowed to grow (van der Zander et al., 2019). To the best of our knowledge, however, it is unknown whether finding an element in $M_{\emptyset,\{i\}}$, for a singleton $\{i\}$ is NP-hard.

C. The Maximum Number of Minimally Invariant Sets

If one does not have a priori knowledge about the graph of the system being analyzed, one can still apply Theorem 5.2 with a correction factor 2^d, as this ensures (with high probability) that no minimally invariant sets are falsely rejected. However, we know that the correction factor is strictly conservative, as there cannot exist 2^d minimally invariant sets in a graph. Thus, correcting for 2^d tests, controls the familywise error rate (FWER) among minimally invariant sets, but increases the risk of falsely accepting a non-invariant set relatively more than what is necessary to control the FWER. Here, we discuss the maximum number of minimally invariant sets that can exist in a graph with d predictor nodes and how a priori knowledge about the sparsity of the graph and the number of interventions can be leveraged to estimate a less strict correction that still controls the FWER.

As minimally invariant sets only contain ancestors of Y (see Proposition 3.3), we only need to consider graphs where Y comes last in a causal ordering. Since d-separation is equivalent to undirected separation in the moralized ancestral graph (Lauritzen, 1996), finding the largest number of minimally invariant sets is equivalent to finding the maximum number of minimal separators in an undirected graph with $d+2$ nodes. It is an open question how many minimal separators exists in a graph with $d+2$ nodes, but it is known that a lower bound for the maximum number of minimal separators is in $\Omega(3^{d/3})$ (Gaspers & Mackenzie, 2015). We therefore propose using a correction factor of $C = 3^{\lceil d/3 \rceil}$ when estimating the set $\hat{S}_{\text{IAS}}$ from Theorem 5.2 if one does not have a priori knowledge of the number of minimally invariant sets in the DAG of the SCM being analyzed. This is a heuristic choice and is not conservative for all graphs.

Theorem 5.2 assumes asymptotic power of the invariance test, but as we can only have a finite amount of data, we will usually not have full power against all non-invariant sets that are not supersets of a minimally invariant set. Therefore, choosing a correction factor that is potentially too low represents a trade-off between error types: if we correct too little, we stand the risk of falsely rejecting a minimally invariant set but not rejecting a superset of it, whereas when correcting too harshly, there is a risk of failing to reject non-invariant sets due to a lack of power.

If one has a priori knowledge of the sparsity or the number of interventions, these can be leveraged to estimate the maximum number of minimally invariant sets using simulation, by the following procedure:

1. For $b = 1, \dots, B$:
   (a) Sample a DAG with d predictor nodes, $N_{\text{interventions}} \sim \mathbb{P}_N$ interventions and $p \sim \mathbb{P}_p$ probability of an edge being present in the graph over (X, Y), such that Y is last in a causal ordering. The measures $\mathbb{P}_N$ and $\mathbb{P}_p$ are distributions representing a priori knowledge. For instance, in a controlled experiment, the researcher may have chosen the number N_0 of interventions. Then, $\mathbb{P}_N$ is a degenerate distribution with $\mathbb{P}_N(N_0) = 1$.
   (b) Compute the set of all minimally invariant sets, e.g., using the `adjustmentSets` algorithm from `dagitty` (Textor et al., 2016).
   (c) Return the number of minimally invariant sets.
2. Return the largest number of minimally sets found in the B repetitions above.

Instead of performing B steps, one can continually update the largest number of minimally invariant sets found so far and end the procedure if the maximum has not updated in a predetermined number of steps, for example.

D. A Finite Sample Algorithm for Computing $\hat{S}_{\text{IAS}}$

In this section, we provide an algorithm for computing the sets $\hat{S}_{\text{IAS}}$ and $\hat{S}^m_{\text{IAS}}$ presented in Theorems 5.2 and 5.5. The algorithm finds minimally invariant sets by searching for invariant sets among sets of increasing size, starting from the empty set. This is done, because the first (correctly) accepted invariant is a minimally invariant set. Furthermore, any set that is a superset of an accepted invariant set, does not need to be tested (as this set cannot be minimal). Tests for invariance can be computationally expensive if one has large amounts of data. Therefore, skipping unnecessary tests offers a significant speedup. In the extreme case, where all singletons are found to be invariant, the algorithm completes in $d+1$ steps, compared to $\sum_{i=0}^{m} \binom{d}{i}$ steps (2^d if $m = d$). This is implemented in lines 8-10 of Algorithm 1.

E. Additional Experiment Details

E.1. Simulation Details for Section 6.1

We sample graphs that satisfy Assumption 2.1 with the additional requirement that $Y \in \text{DE}_Y$ by the following procedure:

1. Sample a DAG $\mathcal{G}$ for the graph of (X, Y) with $d+1$ nodes, for $d \in \{4, 6, \dots, 20\} \cup \{100, 1{,}000\}$, and choose Y to be a node (chosen uniformly at random) that is not a root node.
2. Add a root node E to $\mathcal{G}$ with $N_{\text{interventions}}$ children that are not Y. When $d \leq 20$, $N_{\text{interventions}} \in \{1, \dots, d\}$ and when $d \geq 100$, $N_{\text{interventions}} \in \{1, \dots, 0.1 \times d\}$ (i.e., we consider interventions on up to ten percent of the predictor nodes).
3. Repeat the first two steps if $Y \notin \text{DE}_E$.

Algorithm 1 An algorithm for computing $\hat{S}_{\text{IAS}}$ from data

input A decision rule ϕ_n for invariance, significance thresholds α_0, α, max size of sets to test m (potentially $m = d$) and data

output The set $\hat{S}_{\text{IAS}}$

Initialize $\widehat{\mathcal{MI}}$ as an empty list.
$PS \leftarrow \{S \subseteq [d] \mid |S| \leq m\}$
if $\phi_n(\emptyset, \alpha_0) = 0$ **then**
  End the procedure and return $\hat{S}_{\text{IAS}} = \emptyset$
end if
Sort PS in increasing order according the set sizes
for $S \in PS$ **do**
  if $S \supsetneq S'$ for any $S' \in \widehat{\mathcal{MI}}$ **then**
    Skip the test of S and go to next iteration of the loop
  else
    Add S to $\widehat{\mathcal{MI}}$ if $\phi_n(S, \alpha) = 0$, else continue
  end if
  if The union of $\widehat{\mathcal{MI}}$ contains all nodes **then**
    Break the loop
  end if
end for
Return $\hat{S}_{\text{IAS}}$ as the union of all sets in $\widehat{\mathcal{MI}}$

E.2. Simulation Details for Section 6.2

We simulate data for the experiment in Section 6.2 (and the additional plots in Appendix E.4) by the following procedure:

1. Sample data from a single graph by the following procedure:
   (a) Sample a random graph $\mathcal{G}$ of size $d + 1$ and sample Y (chosen uniformly at random) as any node that is not a root node in this graph.
   (b) Sample coefficients, $\beta_{i \to j}$, for all edges $(i \to j)$ in $\mathcal{G}$ from $U((-2, 0.5) \cup (0.5, 2))$ independently.
   (c) Add a node E with no incoming edges and $N_{\text{interventions}}$ children, none of which are Y. When $d = 6$, we set $N_{\text{interventions}} = 1$ and when $d = 100$, we sample $N_{\text{interventions}}$ uniformly from $\{1, \dots, 10\}$.
   (d) If Y is not a descendant of E, repeat steps (a), (b) and (c) until a graph where $Y \in \text{DE}_E$ is obtained.
   (e) For $n \in \{10^2, 10^3, 10^4, 10^5\}$:
      i. Draw 50 datasets of size n from an SCM with graph $\mathcal{G}$ and coefficients $\beta_{i \to j}$ and with i.i.d. $N(0, 1)$ noise innovations. The environment variable, E, is sampled independently from a Bernoulli distribution with probability parameter $p = 0.5$, corresponding to (roughly) half the data being observational and half the data interventional. The data are generated by looping through a causal ordering of (X, Y), starting at the bottom, and standardizing a node by its own empirical standard deviation before generating children of that node; that is, a node X_j is first generated from PA_j and then standardized before generating any node in CH_j. If X_j is intervened on, we standardize it prior to the intervention.
      ii. For each sampled dataset, apply IAS and ICP. Record the Jaccard similarities between IAS and AN_Y and between ICP and AN_Y, and record whether or not is was a subset of AN_Y and whether it was empty.
      iii. Estimate the quantity plotted (average Jaccard similarity in Figure 4 or probability of $\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y$ or $\hat{S}_{\text{IAS}} = \emptyset$ in Figure 7) from the 50 simulated datasets.
   (f) Return the estimated quantities from the previous step.
2. Repeat the above 100 times and save the results in a data-frame.

E.3. Analysis of the Choice of C in Section 6.2

We have repeated the simulation with $d = 6$ from Section 6.2 but with a correction factor of $C = 2^6$, as suggested by Theorem 5.2 instead of the heuristic correction factor of $C = 9$ suggested in Appendix C. Figure 6 shows the results. We

see that the results are almost identical to those presented in Figure 4. Thus, in the scenario considered here, there is no change in the performance of $\hat{S}_{\text{IAS}}$ (as measured by Jaccard similarity) between using a correction factor of $C = 2^6$ and a correction factor of $C = 3^{\lceil 6/3 \rceil} = 9$. In larger graphs, it is likely that there is a more pronounced difference. E.g., at $d = 10$, the strictly conservative correction factor suggested by Theorem 5.2 is $2^{10} = 1024$, whereas the correction factor suggested in Appendix C is only $3^{\lceil 10/3 \rceil} = 3^4 = 81$, and at $d = 20$ the two are $2^{20} = 1{,}048{,}576$ and $3^{\lceil 20/3 \rceil} = 3^7 = 2187$.

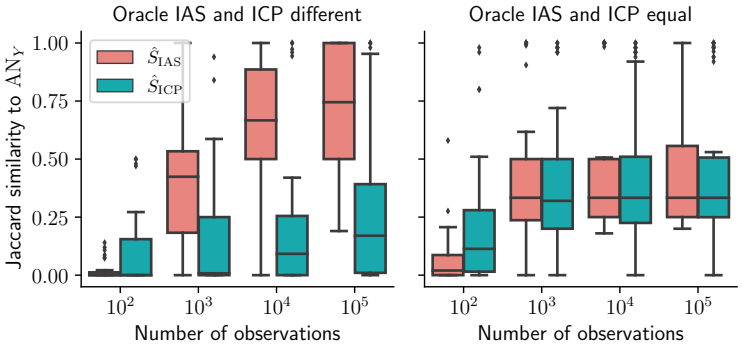


Figure 6. The same figure as in Figure 4, but with a correction factor of $C = 2^6 = 64$ instead of $C = 3^{\lceil 6/3 \rceil} = 9$. Only $d = 6$ shown here, as the correction factor for $d = 100$ is unchanged. Here, the guarantees of Theorem 5.2 are not violated by a potentially too small correction factor, and the results are near identical to those given in Figure 4 using a milder correction factor.

E.4. Analysis of the Choice of α_0 in Section 6.2

Here, we investigate the quantities $\mathbb{P}(\hat{S}_{\text{IAS}} \subseteq \text{AN}_Y)$, $\mathbb{P}(\hat{S}^1_{\text{IAS}} \subseteq \text{AN}_Y)$, $\mathbb{P}(\hat{S}_{\text{IAS}} = \emptyset)$ and $\mathbb{P}(\hat{S}^1_{\text{IAS}} = \emptyset)$ using the same simulation setup as described in Section 6.2. Furthermore, we also ran the simulations for values $\alpha_0 = \alpha$ (testing all hypotheses at the same level), $\alpha_0 = 10^{-6}$ (conservative, see Remark 5.3) as in Section 6.2 and $\alpha_0 = 10^{-12}$ (very conservative). The results for $\alpha = 10^{-6}$ (shown in Figure 7) were recorded in the same simulations that produced the output for Figure 4. For $\alpha_0 \in \{\alpha, 10^{-12}\}$ (shown in Figure 8 and Figure 9, respectively) we only simulated up to 10,000 observations, to keep computation time low.

Generally, we find that the probability of IAS being a subset of the ancestors seems to generally hold well and even more so with large sample sizes. (see Figures 7 to 9), in line with Theorem 5.2. When given 100,000 observations, the probability of IAS being a subset of ancestors is roughly equal to one for almost all SCMs, although there are a few SCMs, where IAS is never a subset of the ancestors (see Figure 7). For $\alpha_0 = 10^{-6}$, the median probability of IAS containing only ancestors is one in all cases, except for $d = 100$ with 1,000 observations – here, the median probability is 87%.

In general, varying α_0 has the effect hypothesized in Remark 5.3: lowering α_0 increases the probability that IAS contains only ancestors, but at the cost of increasing the probability that it is empty (see Figures 7 to 9). For instance, the median probability of IAS being a subset of ancestors when $\alpha_0 = 10^{-12}$ is one for all sample sizes, but the output is always empty when there are 100 observations and empty roughly half the time even at 1,000 observations when $d = 100$ (see Figure 9). In contrast, not testing the empty set at a reduced level, means that the output of IAS is rarely empty, but the probability of IAS containing only ancestors decreases. Still, even with $\alpha_0 = \alpha$, the median probability of IAS containing only ancestors was never lower than 80% (see Figure 8). Thus, choosing α_0 means choosing a trade-off between finding more ancestor-candidates, versus more of them being false positives.

E.5. Analysis of the strength of inverventions in Section 6.2

Here, we repeat the $d = 6$ simulations from Section 6.2 with a reduced strength of the environment to investigate the performance of IAS under weaker interventions. We sample from the same SCMs as sampled in Section 6.2, but reduce the strength of the interventions to be 0.5 instead of 1. That is, the observational distributions are the same as in Section 6.2, but interventions to a node X_j are here half as strong as in Section 6.2.

The Jaccard similarity between $\hat{S}_{\text{IAS}}$ and AN_Y is generally lower than what we found in Figure 4 (see Figure 10). This is likely due to having lower power to detect non-invariance, which has two implications. First, lower power means that we

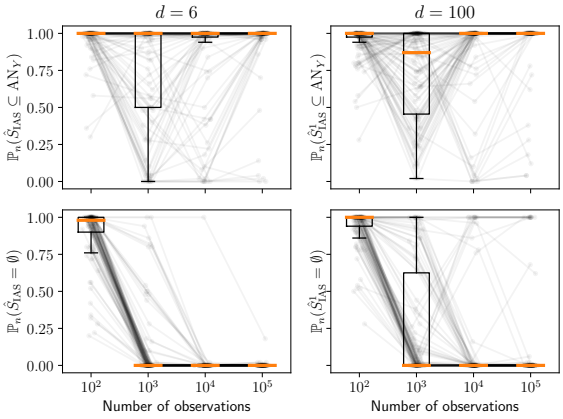


Figure 7. The empirical probabilities of recovering a subset of AN_Y (top row) and recovering an empty set (bottom row), when testing the empty set for invariance at level $\alpha_0 = 10^{-6}$. Generally, our methods seem to hold level well, especially when sample sizes are large. When the sample size is small, the output is often the empty set. When $d = 6$, we estimate $\hat{S}_{\mathrm{IAS}}$ (left column) and when $d = 100$, we estimate $\hat{S}^1_{\mathrm{IAS}}$ (right column). The results here are from the simulations that also produced Figure 4. Medians are displayed as orange lines through each boxplot. Each point represents the probability that the output set is ancestral (resp. empty) for a randomly selected SCM, as estimated by repeatedly sampling data from the same SCM for every $n \in \{10^2, 10^3, 10^4, 10^5\}$. Observations from the same SCM are connected by a line. Each figure contains data from 100 randomly drawn SCMs. Points have been perturbed slightly along the x-axis to improve readability.

may fail to reject the empty set, meaning that we output nothing. Then, the Jaccard similarity between $\hat{S}_{\mathrm{IAS}}$ and AN_Y is zero. Second, it may be that we correctly reject the empty set, but fail to reject another non-invariant set which is not an ancestor of Y which is then potentially included in the output. Then, the $\hat{S}_{\mathrm{IAS}}$ and AN_Y is lower, because we increase the number of false findings.

We find that the probability that $\hat{S}_{\mathrm{IAS}}$ is a subset of ancestors is generally unchanged for the lower intervention strength, but the probability of $\hat{S}_{\mathrm{IAS}}$ generally increases for small sample sizes (see Table 1). This indicates that IAS does not make more mistakes under the weaker interventions, but it is more often uninformative. We see also that in both settings, $\hat{S}_{\mathrm{IAS}}$ is empty more often than $\hat{S}_{\mathrm{ICP}}$ for low sample sizes, but less often for larger samples (see Table 1). This is likely because IAS tests the empty set at a much lower level than ICP does (10^{-6} compared to 0.05). Thus, IAS requires more power to find anything, but once it has sufficient power, it finds more than ICP (see also Figure 10). The median probability of ICP returning a subset of the ancestors was always at least 95% (not shown).

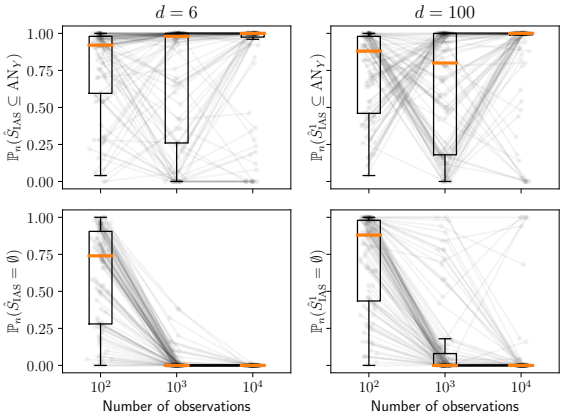


Figure 8. The same figure as Figure 7, but with $\alpha_0 = \alpha = 0.05$ and $n \in \{10^2, 10^3, 10^4\}$. Testing the empty set at the non-conservative level $\alpha_0 = \alpha$ means that the empty set is output less often for small sample sizes, but decreases the probability that the output is a subset of ancestors. Thus, we find more ancestor-candidates, but make more mistakes when $\alpha_0 = \alpha$. However, the median probability of the output being a subset of ancestors is at least 80% in all configurations.

Table 1. Summary of the quantities $\mathbb{P}(\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y)$, $\mathbb{P}(\hat{S}_{\mathrm{IAS}} = \emptyset)$ and $\mathbb{P}(\hat{S}_{\mathrm{ICP}} = \emptyset)$ for weak and strong do-interventions (strength 0.5 and 1, respectively) when $d = 6$. Numbers not in parentheses are means, numbers in parentheses are medians. The level is generally unchanged when the environments have a weaker effect, but the power is lower, in the sense that the empty set is output more often.

		$\mathbb{P}(\hat{S}_{\mathrm{IAS}} \subseteq \mathrm{AN}_Y)$	$\mathbb{P}(\hat{S}_{\mathrm{IAS}} = \emptyset)$	$\mathbb{P}(\hat{S}_{\mathrm{ICP}} = \emptyset)$
Strong interventions	$n = 100$	96.6% (100%)	89.6% (98%)	52.3% (52%)
	$n = 1{,}000$	75.7% (100%)	10.0% (0%)	30.4% (14%)
	$n = 10{,}000$	83.7% (100%)	1.0% (0%)	24.9% (10%)
	$n = 100{,}000$	93.8% (100%)	0.2% (0%)	22.9% (10%)
Weak interventions	$n = 100$	99.3% (100%)	98.7% (100%)	72.0% (84%)
	$n = 1{,}000$	81.1% (100%)	40.2% (26%)	36.9% (24%)
	$n = 10{,}000$	80.8% (100%)	1.7% (0%)	27.5% (15%)
	$n = 100{,}000$	92.6% (100%)	1.1% (0%)	24.8% (14%)

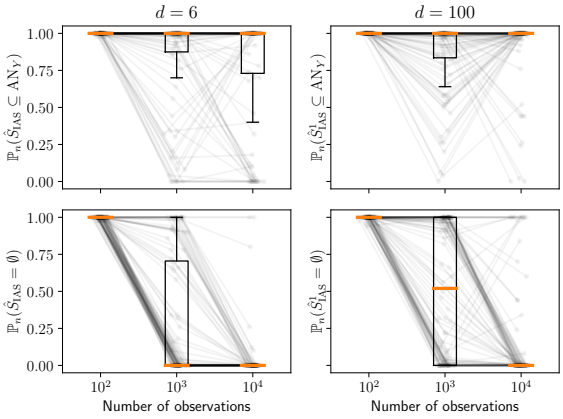


Figure 9. The same figure as Figure 7, but with $\alpha_0 = 10^{-12}$ and $n \in \{10^2, 10^3, 10^4\}$. Testing the empty set at at very conservative level $\alpha_0 = 10^{-12}$ means that the empty set is output more often (for one hundred observations, we only find the empty set), but increases the probability that the output is a subset of ancestors. Thus, testing at a very conservative level $\alpha_0 = 10^{-12}$ means that we do not make many mistakes, but the output is often non-informative.

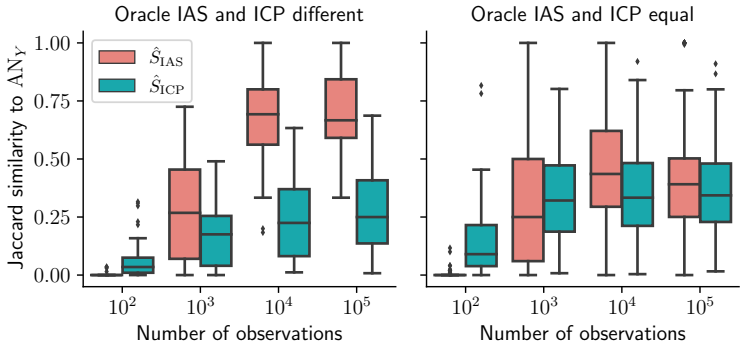


Figure 10. The same figure as the one presented in Figure 4, but with weaker environments (do-interventions of strength 0.5 compared to 1 in Figure 4). Generally, IAS performs the same for weaker interventions as for strong interventions, when there are more than 10,000 observations. Graphs represented in each boxplot: 42 (left), 58 (right).

E.6. Analysis of the Choice of q_{TB} in Section 6.3

In this section, we analyze the effect of changing the cut-off q_{TB} that determines when a gene pair is considered a true positive in Section 6.3. For the results in the main paper, we use $q_{TB} = 1\%$, meaning that the pair $(\texttt{gene}_X, \texttt{gene}_Y)$ is considered a true positive if the value of $\texttt{gene}_Y$ when intervening on $\texttt{gene}_X$ is outside of the 0.01- and 0.99-quantiles of $\texttt{gene}_Y$ in the observational distribution. In Figure 11, we plot the true positive rates for several other choices of q_{TB}. We compare to the true positive rate of random guessing, which also increases if the criterion becomes easier to satisfy. We observe that the choice of q_{TB} does not substantially change the excess true positive rate of our method compared to random guessing. This indicates that while the true positives in this experiments are inferred from data, the conclusions drawn in Figure 5 are robust with respect to some modelling choices of q_{TB}.

E.7. Learning causal ancestors by estimating the I-MEC

In this section, we repeat the experiments performed in Section 6.2, this time including a procedure (here denoted $\text{IAS}_{\text{est. graph}}$), where we perform the following steps.

1. Estimate a member graph of the I-MEC and the location of the intervention sites using Unknown-Target Interventional Greedy Sparsest Permutation (UT-IGSP) (Squires et al., 2020) using the implemention from the Python package

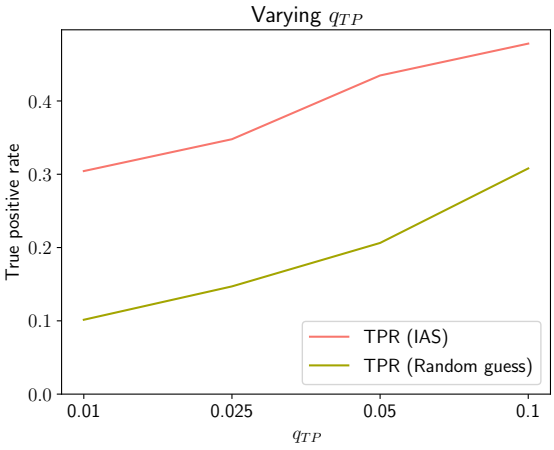


Figure 11. True positive rates (TPRs) for the gene experiment in Section 6.3. q_{TB} specifies the quantile in the observed distribution that an intervention effect has to exceed to be considered a true positive. While the TPR increases for our method when q_{TB} is increased, the TPR of random guessing increases comparably. This validates that changing the definition of true positives in this experiment by choosing a different q_{TB} does not change the conclusion of the experiment substantially.

CausalDAG.[9]

2. Apply the oracle algorithm described in Section 4 to the estimated graph to obtain an estimate of $\mathcal{MI}$.
3. Output the union of all sets in the estimate of $\mathcal{MI}$.

The results for the low-dimensional experiment are displayed in Figure 12 and the results for the high-dimensional experiment are displayed in Table 2. Here, we see that $\text{IAS}_{\text{est. graph}}$ generally performs well (as measured by Jaccard similarity) in the low-dimensional setting ($d = 6$), and even better than IAS for sample sizes $N \leq 10^3$, but is slightly outperformed by IAS for larger sample sizes. However, in the high-dimensional setting ($d = 100$), we observe that $\text{IAS}_{\text{est. graph}}$ fails to hold level and identifies only very few ancestors (see Table 2). We hypothesize that the poor performance of $\text{IAS}_{\text{est. graph}}$ in the high-dimensional setting is due to $\text{IAS}_{\text{est. graph}}$ attempting to solve a more difficult task than IAS. $\text{IAS}_{\text{est. graph}}$ first estimates a full graph (here using UT-IGSP), even though only a subgraph of the full graph is of relevance in this scenario. In addition, UT-IGSP aims to estimate the site of the unknown interventions. In contrast, IAS only needs to identify nodes that are capable of blocking all paths between two variables, and does not need to know the site of the interventions.

	$d = 100, N = 10^3$		$d = 100, N = 10^4$		$d = 100, N = 10^5$	
	IAS	$\text{IAS}_{\text{est. graph}}$	IAS	$\text{IAS}_{\text{est. graph}}$	IAS	$\text{IAS}_{\text{est. graph}}$
$\mathbb{P}(S. \subseteq \text{AN}_Y)$	84.64%	15.30%	94.04%	14.92%	94.72%	14.74%
$\mathbb{P}(S. = \emptyset)$	51.96%	12.32%	12.72%	11.84%	6.98%	11.42%
$J(S., \text{AN}_Y)$	0.19	0.10	0.33	0.10	0.35	0.11

Table 2. Identifying ancestors by first estimating the I-MEC of the underlying DAG and then applying the oracle algorithm of Section 4 fails to hold level and identifies fewer ancestors than applying IAS, when in a high-dimensional setting.

[9] Available at https://github.com/uhlerlab/causaldag.

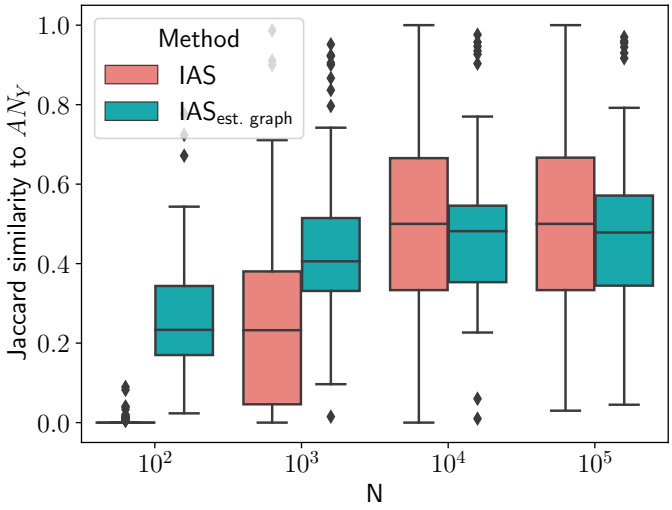


Figure 12. Comparison between the finite sample output of IAS and the procedure described in Appendix E.7, in the low-dimensional case. Generally, these procedures have similar performance, although IAS performs worse for small sample sizes but slightly better for high sample sizes.

4.2 Separating parents from non-parental ancestors

We showed in Paper **C** that IAS outputs non-empty, invariant subsets of ancestors of Y with high probability. The output S_{IAS} of IAS is desirable because 1) it can be used to generate invariant predictions of Y, and 2) it yields information about the potential causes of Y. However, the lack of information about which elements of S_{IAS} are parents and which are non-parental ancestors of Y can lead to some redundancy. For example, in the graph $E \to X_1 \to X_2 \to Y$, IAS outputs $\{X_1, X_2\}$. Without knowing the ground truth DAG behind this learned set of ancestors, we may – for example – decide to make interventions on both X_1 and X_2. However, under atomic interventions, intervening on X_1 and X_2 is equivalent to intervening only on X_1. While intervening on both predictors is not worse than intervening on only one, it may be expensive, time-consuming, or difficult to make multiple interventions. It is therefore valuable to further separate S_{IAS} into parents and non-parental ancestors of Y.

We describe in this section a data-driven way to separate S_{IAS} into parents and non-parental ancestors and prove in Theorem 4.1 that this procedure succeeds with high probability in the limit. Additionally, we perform simulation experiments to show that the separation into parents and non-parents succeeds with high probability in finite samples as well.

To prove Theorem 4.1, we first show that the non-parental ancestors in S_{IAS} are independent of Y given the parents in S_{IAS}.

Lemma 4.1. *Let* $S_0 \coloneqq S_{\text{IAS}} \setminus \text{PA}_Y$ *and* $S \coloneqq S_{\text{IAS}} \cap \text{PA}_Y$ *and assume that* $S_0 \neq \emptyset$*. It holds that*

$$S_0 \perp_d Y \mid S.$$

Proof. Let $i \in S_0$ and denote by $\mathcal{G}$ the graph under consideration. We show that $i \perp_d Y \mid S$, which implies the desired result. Assume for contradiction that there exists a path ϵ between i and Y that is open conditionally on S.

Case 1 – ϵ does not contain a collider: If ϵ is a directed path from Y to i, there is a cycle in $\mathcal{G}$, which contradicts the assumption of acyclicity. If ϵ is either directed from i to Y or has the form $Y \leftarrow \cdots \to i$, there is a parent of Y that: 1) lies on ϵ, 2) is not in S, and 3) is contained in a minimally invariant set. Thus, we have reached a contradiction, as S_{IAS} is the union of all minimally invariant sets.

Case 2 – ϵ contains a collider c: In order for ϵ to be open given S, either the collider itself or a descendant of it must be in S. If c is a descendant of Y, this implies the existence of a cycle (because $S_{\text{IAS}} \subseteq \text{AN}_Y$), contradicting acyclicity of $\mathcal{G}$. If c is a non-descendant of Y, there is a parent of Y between Y and c which is contained in a minimally invariant set but is not in S, contradicting that S_{IAS} is the union of all minimally invariant sets.

Having reached a contradiction in all possible cases, we conclude that $i \perp_d Y \mid S$, and therefore $S_0 \perp_d Y \mid S$. □

Below, we consider the case where Y is linear in all its causes and has additive Gaussian noise.[2] That is, we assume that Y is structurally generated as

$$Y \coloneqq \sum_{i \in \mathrm{PA}_Y} \beta_i X_i + N_Y, \quad N_Y \sim N(0, \sigma^2).$$

Let $\mathcal{D} = (X_i, E_i, Y_i)_{i=1}^N$ be i.i.d. observations of (X, E, Y).[3] The procedure is as follows:

1. Use $\mathcal{D}$ to learn an estimate of S_{IAS}.

2. Regress Y onto $X_{\hat{S}_{\mathrm{IAS}}}$ and test the hypotheses $H_{0,i} : \hat{\beta}_i = 0$ for all $i \in \hat{S}_{\mathrm{IAS}}$. Denote by ϕ_i a decision rule for $H_{0,i}$ being rejected ($\phi_i = 1$) using any method that controls the FWER (e.g., Bonferroni corrected p-values from T-tests).

3. Split $\hat{S}_{\mathrm{IAS}}$ according to the rejected and non-rejected hypotheses:

$$\hat{S}_{\mathrm{IAS}}^{\mathrm{PA}} \coloneqq \bigcup_{i \in \hat{S}_{\mathrm{IAS}}\colon\ \phi_i = 1} \{i\}$$

and

$$\hat{S}_{\mathrm{IAS}}^{\neg \mathrm{PA}} \coloneqq \bigcup_{i \in \hat{S}_{\mathrm{IAS}}\colon\ \phi_i = 0} \{i\}.$$

Below, we show that $\hat{S}_{\mathrm{IAS}}^{\mathrm{PA}}$ (resp. $\hat{S}_{\mathrm{IAS}}^{\neg \mathrm{PA}}$) is a subset of parents of Y (resp. non-parental ancestors of Y) with high probability if the individual hypotheses $H_{0,i}$ can be tested with asymptotic power.

Theorem 4.1. *Assume for all $i \in S_{\mathrm{IAS}} \cap \mathrm{PA}_Y$ that ϕ_i has (uniform or pointwise) asymptotic power to reject H_i. Then*

$$\lim_{n \to \infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}}^{\mathrm{PA}} \subseteq \mathrm{PA}_Y) \geq (1 - \alpha)^2$$

and

$$\lim_{n \to \infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}}^{\neg \mathrm{PA}} \subseteq \mathrm{AN}_Y \setminus \mathrm{PA}_Y) \geq (1 - \alpha)^2.$$

[2] Similar arguments apply to the general case of additive noise models if the regression functions are consistent minimizers of the $\mathcal{L}^2$ loss.

[3] This is a slight abuse of notation: recall that $X = (X_1, \ldots, X_d)$. Here, the subscripts refer to coordinate projections of the d-valued random variable. In $\mathcal{D} = (X_i, E_i, Y_i)_{i=1}^N$, the subscript i refers to an observation of all d marginals – not a coordinate of X.

Proof. Let $\Omega \coloneqq \hat{S}_{\mathrm{IAS}} \setminus \mathrm{PA}_Y$ and $\Theta \coloneqq \hat{S}_{\mathrm{IAS}} \cap \mathrm{PA}_Y$. Then:

$$\begin{aligned}\mathbb{P}(\hat{S}_{\mathrm{IAS}}^{\mathrm{PA}} \subseteq \mathrm{PA}_Y) \geq & \mathbb{P}\left((\hat{S}_{\mathrm{IAS}} = S_{\mathrm{IAS}}) \cap \bigcap_{i \in \Omega}(\phi_i = 0) \cap \bigcap_{i \in \Theta}(\phi_i = 1)\right) \\ = & \mathbb{P}\left(\bigcap_{i \in \Omega}(\phi_i = 0) \cap \bigcap_{i \in \Theta}(\phi_i = 1) \mid \hat{S}_{\mathrm{IAS}} = S_{\mathrm{IAS}}\right) \\ & \times \mathbb{P}(\hat{S}_{\mathrm{IAS}} = S_{\mathrm{IAS}})\end{aligned}$$

From Paper **C**, Theorem 5.2, we know that the quantity $\mathbb{P}(\hat{S}_{\mathrm{IAS}} = S_{\mathrm{IAS}})$ can be asymptotically controlled at level $1-\alpha$. We then consider the remaining term. Given that $\hat{S}_{\mathrm{IAS}} = S_{\mathrm{IAS}}$, Lemma 4.1 implies that every hypothesis in Ω is true and every hypothesis in Θ is false. Thus, conditionally on $\hat{S}_{\mathrm{IAS}} = S_{\mathrm{IAS}}$, the term simply states that we reject all false hypotheses and fail to reject all true hypotheses. By the assumption of asymptotic power, all false hypotheses are rejected with probability one in the limit. Furthermore, we have constructed the tests ϕ_i, such that the FWER is controlled at level α. Thus, the remaining term is asymptotically controlled at level $1-\alpha$. In summary, we conclude that

$$\lim_{n \to \infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}}^{\mathrm{PA}} \subseteq \mathrm{PA}_Y) \geq (1-\alpha)^2.$$

One can apply the same argument to show that $\lim_{n\to\infty} \mathbb{P}(\hat{S}_{\mathrm{IAS}}^{\neg \mathrm{PA}} \subseteq \mathrm{AN}_Y \setminus \mathrm{PA}_Y) \geq (1-\alpha)^2$. □

We verify Theorem 4.1 by simulation. We consider the same simulation setup as in Paper **C**, Section 6.2. This time, we follow up the estimation of $\hat{S}_{\mathrm{IAS}}$ by splitting it into $\hat{S}_{\mathrm{IAS}}^{\mathrm{PA}}$ and $\hat{S}_{\mathrm{IAS}}^{\neg \mathrm{PA}}$ as described above. We sample 100 graphs and corresponding coefficients, and for each $N \in \{10^2, \ldots, 10^5\}$, we sample 50 data sets of N observations. For each obtained data set, we estimate S_{IAS} and:

1. Linearly regress Y onto the learned set $\hat{S}_{\mathrm{IAS}}$.
2. Test each hypothesis $H_{0,i} : \beta_i = 0$, $i \in \hat{S}_{\mathrm{IAS}}$, using a two-sided T-test.
3. Reject $H_{0,i}$ if the p-value from the T-test is below $0.05/|\hat{S}_{\mathrm{IAS}}|$.
4. Construct $\hat{S}_{\mathrm{IAS}}^{\mathrm{PA}}$ and $\hat{S}_{\mathrm{IAS}}^{\neg \mathrm{PA}}$.
5. Record:
   - Whether $\hat{S}_{\mathrm{IAS}}^{\mathrm{PA}} \subseteq \mathrm{PA}_Y$ and $\hat{S}_{\mathrm{IAS}}^{\neg \mathrm{PA}} \subseteq \mathrm{AN}_Y \setminus \mathrm{PA}_Y$.
   - The Jaccard similarity of $\hat{S}_{\mathrm{IAS}}^{\mathrm{PA}}$ to the set of discoverable parents $S_{\mathrm{IAS}} \cap \mathrm{PA}_Y$, and the Jaccard similarity of $\hat{S}_{\mathrm{ICP}}$ to $S_{\mathrm{IAS}} \cap \mathrm{PA}_Y$.
   - The size of the sets $\hat{S}_{\mathrm{IAS}}^{\mathrm{PA}}$ and $\hat{S}_{\mathrm{ICP}}$.

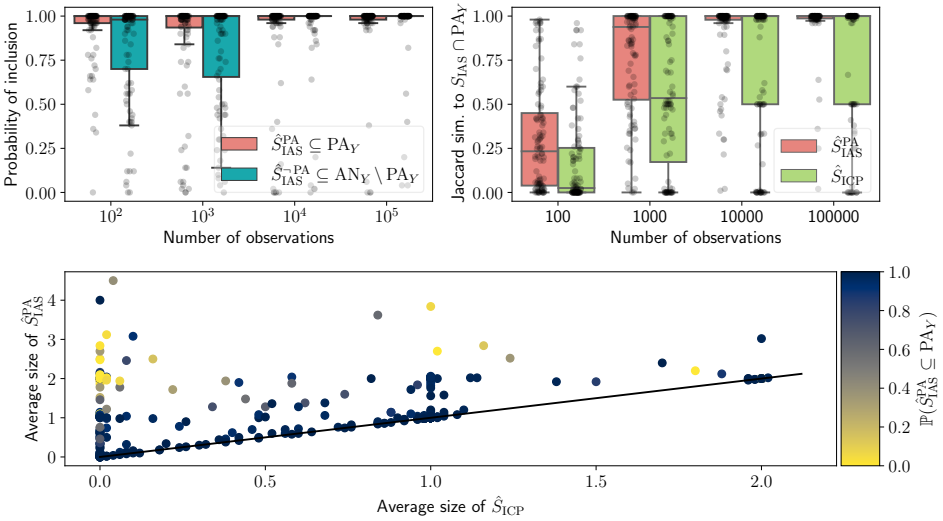


Figure 4.1: Results of the simulation described in Section 4.2. Top left: The empirical probabilities that the output sets $\hat{S}^{\mathrm{PA}}_{\mathrm{IAS}}$ and $\hat{S}^{\neg\,\mathrm{PA}}_{\mathrm{IAS}}$ contain only parents of Y and non-parental ancestors of Y, respectively. Top right: The Jaccard similarity between the set of discoverable parents $S_{\mathrm{IAS}} \cap \mathrm{PA}_Y$ and $\hat{S}_{\mathrm{ICP}}$ and $\hat{S}^{\mathrm{PA}}_{\mathrm{IAS}}$, respectively. Bottom: The average size of $\hat{S}_{\mathrm{ICP}}$ versus the average size of $\hat{S}_{\mathrm{ICP}}$. The black line is the identity line. Each point represents the average of 50 estimates on the same graph with the same number of observations. Points are colored by the empirical probability that $\hat{S}^{\mathrm{PA}}_{\mathrm{IAS}} \subseteq \mathrm{PA}$. In the top graphs, points have been jittered slightly along the x-axis to increase readability.

The results are summarized in Figure 4.1. From this, we see three things. First, the output sets $\hat{S}^{\mathrm{PA}}_{\mathrm{IAS}}$ and $\hat{S}^{\neg\,\mathrm{PA}}_{\mathrm{IAS}}$ generally hold level even for relatively small sample sizes. Second, $\hat{S}^{\mathrm{PA}}_{\mathrm{IAS}}$ generally has a higher Jaccard similarity to the set of discoverable parents $S_{\mathrm{IAS}} \cap \mathrm{PA}_Y$ than $\hat{S}_{\mathrm{ICP}}$ does. Third, in cases where ICP finds no candidate parents, $\hat{S}^{\mathrm{PA}}_{\mathrm{IAS}}$ appears sometimes to be overly volatile. That is, $\hat{S}^{\mathrm{PA}}_{\mathrm{IAS}}$ will output a large set of predictors, but this set also contains non-parents. This is likely to happen in the cases where we do not have sufficient power to reject non-invariant sets, in which case $\hat{S}_{\mathrm{IAS}}$ potentially contains descendants of Y. In other words: when everything looks invariant (but is not invariant), IAS outputs everything and ICP outputs nothing.

As shown in Paper **C**, Proposition 3.4, the oracle output of ICP equals that of IAS if and only if the oracle output of ICP is invariant. In Figure 4.2, we have repeated a modified version of the experiment from Figure 4.1. This time, we sample 100 graphs in which the oracle output of ICP is not invariant (i.e., $S_{\mathrm{ICP}} \subsetneq S_{\mathrm{IAS}}$). As expected, we find that $\hat{S}^{\mathrm{PA}}_{\mathrm{IAS}}$ still has a high Jaccard similarity to the set of discoverable parents, but that $\hat{S}_{\mathrm{ICP}}$ generally has low

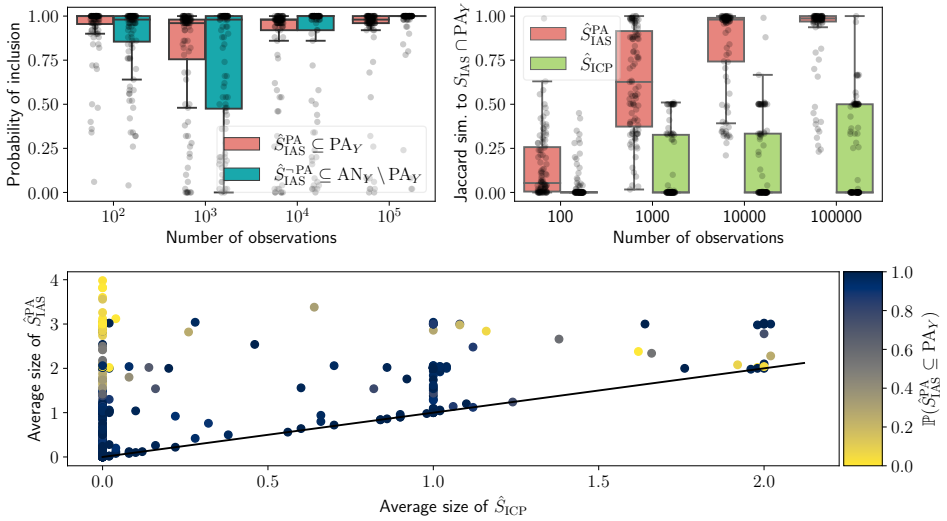


Figure 4.2: Same experiment and figure as in Figure 4.1, with the additional requirement that $S_{\text{ICP}} \subsetneq S_{\text{IAS}}$ in all graphs.

similarity to the discoverable parents. The levels of $\hat{S}^{\text{PA}}_{\text{IAS}}$ and $\hat{S}^{\neg\text{PA}}_{\text{IAS}}$ are largely unchanged.

In summary, the results of this section demonstrate that the output of IAS can be post hoc separated into parents and non-parental ancestors of Y while retaining asymptotic guarantees. This yields additional information about the causal structure of the subgraph induced by Y and S_{IAS}, and facilitates the discovery of more parents than by applying ICP. Like in Paper **C**, this comes at the cost of all guarantees being asymptotic and dependent on several tests having asymptotic power.

Bibliography

JJ Allaire, Romain Francois, Kevin Ushey, Gregory Vandenbrouck, Marcus Geelnard, and Intel. *RcppParallel: Parallel Programming Tools for 'Rcpp'*, 2022. URL `https://CRAN.R-project.org/package=RcppParallel`. R package version 5.1.5.

Yoav Benjamini and Yosef Hochberg. 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, 1995.

Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83, 2000.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Antonio Colangelo, Marco Scarsini, and Moshe Shaked. Some notions of multivariate positive dependence. *Insurance: Mathematics and Economics*, 37 (1):13–26, 2005.

Vanessa Didelez. Graphical models for composable finite markov processes. *Scandinavian Journal of Statistics*, 34(1):169–185, 2007.

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

Vanessa Didelez. Causal reasoning for events in continuous time: A decision-theoretic approach. In *ACI@ UAI*, pages 40–45, 2015.

Edgar Dobriban. Fast closed testing for exchangeable local tests. *Biometrika*, 107(3):761–768, 2020.

David Donoho and Jiashun Jin. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, 30(1):1–25, 2015.

David Donoho and Jiasun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of statistics*, 32(3):962–994, 2004. ISSN 0090-5364.

Frank Dudbridge and Bobby PC Koeleman. Rank truncated product of p-values, with application to genomewide association scans. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 25(4):360–366, 2003.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

Dirk Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. Springer, New York, 2013. doi: 10.1007/978-1-4614-6868-4. ISBN 978-1-4614-6867-7.

Dirk Eddelbuettel and James Joseph Balamuta. Extending extitR with extitC++: A Brief Introduction to extitRcpp. *The American Statistician*, 72 (1):28–36, 2018. doi: 10.1080/00031305.2017.1375990.

Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011. doi: 10.18637/jss.v040.i08.

Shaun Fallat, Steffen Lauritzen, Kayvan Sadeghi, Caroline Uhler, Nanny Wermuth, and Piotr Zwiernik. Total positivity in markov structures. *The Annals of Statistics*, pages 1152–1184, 2017.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27 (8):861–874, 2006.

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

Quentin Giai Gianetto, Florence Combes, Claire Ramus, Christophe Bruley, Yohann Couté, and Thomas Burger. *cp4p: Calibration Plot for Proteomics*, 2019. URL `https://CRAN.R-project.org/package=cp4p`. R package version 0.3.6.

Jelle J Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

Jelle J Goeman, Rosa J Meijer, Thijmen JP Krebs, and Aldo Solari. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106(4):841–856, 2019.

Jelle J Goeman, Jesse Hemerik, and Aldo Solari. Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics*, 49(2):1218–1238, 2021.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Eugene Grechanovsky and Yosef Hochberg. Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference*, 76 (1-2):79–91, 1999.

Gerhard Hommel and T Hoffmann. Controlled uncertainty. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, pages 154–161. Springer, 1988.

Hongmei Jiang and RW Doerge. Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer informatics*, 6: 117693510800600001, 2008.

Edward L Korn, James F Troendle, Lisa M McShane, and Richard Simon. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398, 2004.

Mette Langaas, Bo Henry Lindqvist, and Egil Ferkingstad. Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):555–572, 2005.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

EL Lehmann. On families of admissible tests. *The Annals of Mathematical Statistics*, pages 97–104, 1947.

EL Lehmann and Joseph P Romano. Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3):1138–1154, 2005.

Yaowu Liu and Jun Xie. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020. doi: 10.1080/01621459.2018.1554485. URL `https://doi.org/10.1080/01621459.2018.1554485`. PMID: 33012899.

Ruth Marcus, Peritz Eric, and K Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3): 655–660, 1976.

Nicolai Meinshausen and John Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics*, 34(1):373–393, 2006.

Phillip B. Mogensen. *TMTI: 'Too Many, Too Improbable' (TMTI) Test Procedures*, 2021. URL `https://CRAN.R-project.org/package=TMTI`. R package version 0.1.0.

Søren Wengel Mogensen. Causal screening in dynamical systems. In *Conference on Uncertainty in Artificial Intelligence*, pages 310–319. PMLR, 2020.

Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1): 539–559, 2020.

Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *UAI*, pages 350–360, 2018.

Dan Nettleton, JT Hwang, Rico A Caldo, and Roger P Wise. Estimating the number of true null hypotheses from a histogram of p values. *Journal of agricultural, biological, and environmental statistics*, 11(3):337–356, 2006.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Jonas Peters, Stefan Bauer, and Niklas Pfister. Causal models for dynamical systems. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 671–690. Association for Computing Machinery, 2022.

Paul J Pockros, Dominque Guyader, Heather Patton, Myron J Tong, Terry Wright, John G McHutchison, and Tze-Chiang Meng. Oral resiquimod in chronic hcv infection: safety and efficacy in 2 placebo-controlled, double-blind phase iia studies. *Journal of hepatology*, 47(2):174–182, 2007.

Stan Pounds and Cheng Cheng. Robust estimation of the false discovery rate. *Bioinformatics*, 22(16):1979–1987, 2006.

Jacob Agerbo Rasmussen, Kasper Rømer Villumsen, Madeleine Ernst, Martin Hansen, Torunn Forberg, Shyam Gopalakrishnan, M Thomas P Gilbert, Anders Miki Bojesen, Karsten Kristiansen, and Morten Tønsberg Limborg. A multi-omics approach unravels metagenomic and metabolic alterations of a probiotic and synbiotic additive in rainbow trout (oncorhynchus mykiss). *Microbiome*, 10(1):1–19, 2022.

Alexander Gilbert Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! causal discovery benchmarks may be easy to game. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=wEOlVzVhMW_`.

Joseph P Romano, Azeem Shaikh, and Michael Wolf. Consonance and the closure method in multiple testing. *The International Journal of Biostatistics*, 7(1), 2011.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019.

Sanat K Sarkar. Some probability inequalities for ordered mtp2 random variables: a proof of the simes conjecture. *Annals of Statistics*, pages 494–504, 1998.

Sanat K Sarkar and Woollcott Smith. Probability inequalities for ordered random variables. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 119–135, 1986.

Tore Schweder. Composable markov processes. *Journal of applied probability*, 7(2):400–410, 1970.

Jonas Seng, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Tearing apart notears: Controlling the graph prediction via variance manipulation. *arXiv preprint arXiv:2206.07195*, 2022.

Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

Eckart Sonnemann and Helmut Finner. Vollständigkeitssätze für multiple testprobleme. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, pages 121–135. Springer, 1988.

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

John D Storey, Jonathan E Taylor, and David Siegmund. 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, 2004.

Jinjin Tian, Xu Chen, Eugene Katsevich, Jelle Goeman, and Aaditya Ramdas. Large-scale simultaneous inference under dependence. *arXiv preprint arXiv:2102.11253*, 2021.

John Wilder Tukey. The problem of multiple comparisons. *Multiple comparisons*, 1953.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

Hong-Qiang Wang, Lindsey K Tuominen, and Chung-Jui Tsai. Slim: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*, 27(2):225–231, 2011.

Daniel J Wilson. The lévy combination test. *arXiv preprint arXiv:2105.01501*, 2021.

Jashin J Wu, David B Huang, and Stephen K Tyring. Resiquimod: a new immune response modifier with potential as a vaccine adjuvant for th1 immune responses. *Antiviral research*, 64(2):79–83, 2004.

Dmitri V Zaykin, Lev A Zhivotovsky, Peter H Westfall, and Bruce S Weir. Truncated product method for combining p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 22(2):170–185, 2002.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.