cemmap

centre for microdata methods and practice

# Testing sign congruence between two parameters

Douglas L. Miller
Francesca Molinari
Jörg Stoye

The Institute for Fiscal Studies
Department of Economics, UCL

**cemmap** working paper CWP14/24

Economic and Social Research Council

# TESTING SIGN CONGRUENCE BETWEEN TWO PARAMETERS


DOUGLAS L. MILLER
Department of Economics, Cornell University

FRANCESCA MOLINARI
Department of Economics, Cornell University

JÖRG STOYE
Department of Economics, Cornell University



We test the null hypothesis that two parameters $(\mu_1, \mu_2)$ have the same sign, assuming that (asymptotically) normal estimators $(\hat{\mu}_1, \hat{\mu}_2)$ are available. Examples of this problem include the analysis of heterogeneous treatment effects, causal interpretation of reduced-form estimands, meta-studies, and mediation analysis. A number of tests were recently proposed. We recommend a test that is simple and rejects more often than many of these recent proposals. Like all other tests in the literature, it is conservative if the truth is near $(0, 0)$ and therefore also biased. To clarify whether these features are avoidable, we also provide a test that is unbiased and has exact size control on the boundary of the null hypothesis, but which has counterintuitive properties and hence we do not recommend. We use the test to improve p-values in Kowalski (2022) from information contained in that paper's main text and to establish statistical significance of some key estimates in Dippel et al. (2021).



Douglas L. Miller: dlm336@cornell.edu

Francesca Molinari: fm72@cornell.edu

Jörg Stoye: stoye@cornell.edu



This version: June of 2024. We thank Hyewon Kim for excellent research assistance and Gregory Cox, David Kaplan, Amanda Kowalski, and audiences at the 2023 Southern Economic Association Annual Meeting and the 2024 Econometric Society North American Winter Meeting for feedback.

# 1. INTRODUCTION

It is common in empirical practice to investigate whether parameters of interest have the same sign. Applications in the analysis of treatment effects loom large. For example, researchers frequently ask whether an average treatment effect has the same sign across different subgroups of the population, as this knowledge helps in determining treatment protocols. Researchers estimating marginal treatment effects or local average treatment effects may similarly investigate whether these effects are heterogeneous across groups. Indeed, Brinch et al. (2017) establish that testing whether the marginal treatment effect is constant across conditioning values of the propensity score can be expressed as a test of whether two parameters have the same sign,[1] and they implement a test of this hypothesis to analyze how family size affects children's educational attainment. Kowalski (2022) carries out a similar test to determine whether "women more likely to receive mammograms experience higher levels of overdiagnosis." Even determining whether instrument validity fails in causal inference with non-compliers may include testing for sign congruence among population parameters (Kim, 2023, Section 2.2). Examples are not confined to the question of treatment effect heterogeneity. The meta-analysis literature investigates whether two or more different studies estimate a treatment effect that has the significantly same sign (see, e.g., Slough and Tyson, 2024). And in the rich literature on causal mediation (see, e.g., MacKinnon et al., 2002, for an overview), one asks whether the sign of the treatment effects in which the overall effect is decomposed are the same. Instances of this testing problem are easy to find also outside the treatment evaluation literature: Liu and Molinari (2024) show that a hypothesis of sign congruence is relevant in the analysis of the properties of an algorithmic fairness-accuracy frontier.

This paper addresses the problem of testing for sign congruence among two parameters as well as some extensions to related hypotheses. Formally, we start with the null hypothesis

$$H_0 : \mu_1 \cdot \mu_2 \geq 0. \tag{1.1}$$

[1] These parameters are the difference in mean outcome for the treated group when a binary instrument takes value 1 and value 0 and the difference in mean outcome for the untreated group when a binary instrument takes value 1 and value 0.

We rediscover and recommend a deceptively simple test, which was proposed earlier by Russek-Cohen and Simon (1993).[2] In the special case (1.1) and if estimators $(\hat{\mu}_1, \hat{\mu}_2)$ are (asymptotically) normal with unit variances and uncorrelated (or positively correlated), the test rejects sign equality at the $5\%$ significance level if $\hat{\mu}_1 \cdot \hat{\mu}_2 < 0$ and furthermore $\min\{|\hat{\mu}_1|, |\hat{\mu}_2|\} > 1.64$. The more general version of the test with unrestricted correlation is only marginally more complicated as it requires one to calibrate the critical value instead of using $1.64$. Nonetheless, the test not only is valid in the sense of controlling size uniformly over all $(\mu_1, \mu_2)$ on the null hypothesis, but is also uniformly more powerful than recently proposed tests that we discuss below. That said, the new test is still quite conservative near a true value of $(0, 0)$ and has limited power against alternatives that are local to that point. We show that this feature can in principle be avoided by providing a test that has uniformly exact size control on the boundary of the null hypothesis and is also unbiased. However, that test rejects for some sample realizations that are arbitrarily close to the null hypothesis. Under an intuitive criterion that excludes such features and for most relevant values of an identified nuisance parameter, our recommended test is *uniformly* most powerful.

A recent literature motivated by some of the above examples addressed the testing problem in (1.1) and put forward different testing procedures than the one we recommend here. Brinch et al. (2017, BMW henceforth) establish that (1.1) can be tested by simultaneously testing the sign of $\mu_1$ and $\mu_2$ and using Bonferroni correction to adjust for multiplicity of tests. They also consider a more general testing problem in which $(\mu_1, \mu_2)$ are hypothesized to lie in the union of two rotation symmetric origin cones whose vertices must be estimated. We generalize our test to that setting and formally show that the test we recommend is more powerful than the one in BMW. Slough and Tyson (2024) use the same Bonferroni adjustment in the context of meta-analysis (though this application frequently calls for generalization to $K > 2$ parameters, which we do not analyze). Kowalski (2022) proposes a heuristic bootstrap test and claims improved power over Bonferroni adjustment. We provide conditions under which these claims are true, verify that they hold in Kowal-

[2]We became aware of this literature reading Kim (2023) after independently deriving a solution to the problem. The test was also foreshadowed by the "LR test" in unpublished sections of Kaplan (2015), a working paper precursor of Kaplan and Zhuo (2021).

ski's (2022) setting, and clarify that this test is not valid without them. Finally, the literature on causal mediation puts forward several testing approaches. Among these, Montoya and Hayes (2017) perform the same bootstrap test as Kowalski (2022).

Kim (2023) proposes a general testing procedure with uniform asymptotic validity that allows to test sign congruence among a finite number $K \geq 2$ of means when the asymptotic correlation matrix of the estimators of these means is unrestricted, along with a refinement to the testing procedure that may yield improved power performance for the case that $K > 2$. In contrast, we only consider the case $K = 2$ with arbitrary correlation.

The rest of this paper contains the following contributions. We restate the testing problem and briefly explain why it is difficult. We next provide our recommended test, which turns out to rediscover a test in Russek-Cohen and Simon (1993). We provide a simple independent proof of its validity and also establish the other, novel claims outlined above. We next clarify when the use of the heuristic bootstrap in Kowalski (2022) and several other applications is justified. Specifically, such validity requires estimators to be (asymptotically) uncorrelated; but in this case, our recommended test is both easier to implement and uniformly more powerful. In general, the heuristic bootstrap test can have arbitrary size. We close with two empirical applications, both of which strengthen the statistical conclusions of published work.

## 2. RECOMMENDED TESTING PROCEDURE

### 2.1. *Explaining the Testing Problem*

We start with a self-contained treatment of the problem of testing the null hypothesis in (1.1), i.e., whether $\mu_1$ and $\mu_2$ have the same sign. We initially assume that we have estimators $(\hat{\mu}_1, \hat{\mu}_2)$ with known covariance matrix and

$$\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right). \tag{2.1}$$

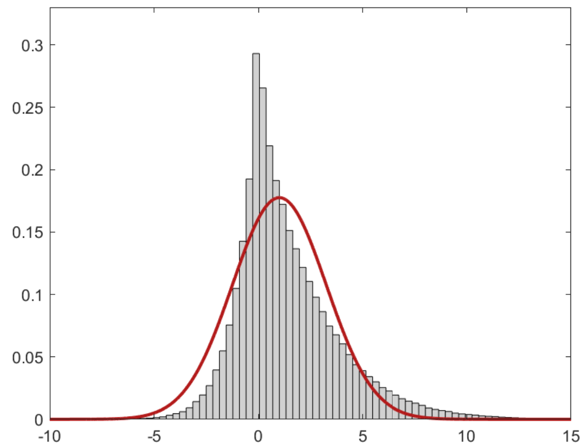


FIGURE 1.—True distribution (histogram) and Delta method approximated distribution (curve) of $\hat{\mu}_1 \cdot \hat{\mu}_2$ if true parameter values are as in (2.2).

We think of this primarily as asymptotic approximation. By entirely standard arguments, all claims that follow can be extended (in an asymptotic sense) to asymptotically normal estimators whose asymptotic variance matrix can be estimated.

This is a special case of a problem that has received attention in statistics; see Datta (1995) and references therein. It is difficult because, near a true parameter value of $(\mu_1, \mu_2) = (0,0)$, the natural "plug-in" test statistic $\hat{t} \equiv \hat{\mu}_1 \cdot \hat{\mu}_2$ will be far from normally distributed, meaning that its distribution cannot be approximated by common bootstrap schemes or the Delta method.[3]

Because the Delta-method works in a pointwise sense except literally at $(0,0)$, one might hope that the problem is "exotic" in the sense of being practically confined to a small region in parameter space. But this is not the case. Figure 1 contrasts the true distribution of $\hat{\mu}_1 \cdot \hat{\mu}_2$

[3]To quickly see the issue, note that, if $\rho = \sigma_1 = \sigma_2 = 1$ and $(\mu_1, \mu_2) = (0,0)$, then the distribution of $\hat{t}$ is $\chi^2_1$. In addition, while we avoid explicit asymptotics, a $\hat{t}$ computed from sample averages would be superconsistent.

The literature on statistics of this type goes back at least to Craig (1936). In the context of causal mediation, Delta method inference was influentially advocated by Sobel (1982), although its invalidity is by now widely recognized in that literature.

with an oracle Delta method approximation (i.e., using true parameter values) when

$$\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 2 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 & .4 \\ .4 & 1 \end{pmatrix} \right). \tag{2.2}$$

Here, one of the true means is two standard deviations away from zero, which one might intuit to be a "safe" distance, but the Delta-method approximation is clearly inaccurate.

An established literature in statistics considers general testing problems of this type and provides remedies that typically rely on pre-testing whether one is close to the point of irregularity and appropriately modifying critical values. Fortunately, much simpler tests are available for the special case of testing the null hypothesis in (1.1). BMW use a combination of Bonferroni and union-intersection arguments to show size control for the following test:

$$\text{Reject } H_0 \text{ if } \hat{\mu}_1 \cdot \hat{\mu}_2 < 0 \text{ and } \min\{|\hat{\mu}_1|/\sigma_1, |\hat{\mu}_2|/\sigma_2\} \geq \Phi(1-\alpha/2).$$

In words, the estimated signs disagree and are individually significant at a level that reflects Bonferroni correction. Kowalski (2022) uses instead a heuristic bootstrap test that can be translated as follows:

$$\text{Reject } H_0 \text{ if } \Pr(\hat{\mu}_1^* \cdot \hat{\mu}_2^* > 0) < \alpha, \text{ where } \begin{pmatrix} \hat{\mu}_1^* \\ \hat{\mu}_2^* \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right).$$

Here, Kowalski (2022) bootstraps the joint distribution of $(\hat{\mu}_1^*, \hat{\mu}_2^*)$. By exactly equating the bootstrap and population distributions of estimation errors, we assume this bootstrap works perfectly, possibly in some asymptotic limit.[4] The test's heuristic motivation is to check whether $0$ is inside a one-sided nonparametric percentile bootstrap confidence interval for $\mu_1 \cdot \mu_2$ that was computed in the spirit of Efron (1979).

[4]As a reminder, the aforementioned use cases involve estimators for which validity of basic bootstrap approximation is essentially equivalent to asymptotic normality; for example, this follows for simple nonparametric or wild bootstrapping of sample moments from Mammen (1992). We therefore do not think of bootstrap validity for the distribution of $(\hat{\mu}_1, \hat{\mu}_2)$ as a meaningful strengthening of our assumptions. Of course,validity of simple percentile bootstrap and related approximations for $\hat{\mu}_1 \cdot \hat{\mu}_2$ does not follow nor, in fact, hold.

It turns out that this problem has been previously considered in the statistics literature on *qualitative interactions*, which proposes methods to test whether treatment effects vary across groups of individuals. In particular, after developing a solution, we found that same solution in Russek-Cohen and Simon (1993, Theorem 6.1), with related results for the case that the estimators are asymptotically uncorrelated in Gail and Simon (1985, Theorem 1). While we provide a self-contained treatment that should be accessible to empirical economists, we do not claim novelty for results reported in this section.

With that said, we recommend the following test:

$$\text{Reject } H_0 \text{ if } \hat{\mu}_1 \cdot \hat{\mu}_2 < 0 \text{ and } \min\{|\hat{\mu}_1|/\sigma_1, |\hat{\mu}_2|/\sigma_2\} \geq c_\alpha, \tag{2.3}$$

where $c_\alpha$ is characterized by

$$\sup_{\mu_1=0,\mu_2\geq 0} \Pr\left(\hat{\mu}_1 \cdot \hat{\mu}_2 \geq c_\alpha\right) = \alpha. \tag{2.4}$$

The value $c_\alpha$ can be computed numerically as a solution to (2.4). In fact, Gail and Simon (1985, Table 1) provide a tabulation for several values of $\alpha$ (and several values of $K$, with $K$ the number of means whose sign congruence one is testing) assuming that their estimators are asymptotically uncorrelated; Russek-Cohen and Simon (1993, Table 2) provide a tabulation for several values of $\rho$ and $\alpha = 0.05$ and $0.01$ for the case of two means. As we show in the proof of Theorem 2.1, $c_\alpha = \Phi(1-\alpha)$ whenever $\rho \geq 0$. Numerically, one can show that $c_\alpha = \Phi(1-\alpha)$ whenever $\rho > -0.75$ (see, e.g., Russek-Cohen and Simon, 1993, Table 2). Notably, $\rho \geq 0$ includes the case where estimates (asymptotically) derive from distinct subsamples and therefore $\rho = 0$. In this latter case, the test simplifies to

$$\text{Reject } H_0 \text{ if } \hat{\mu}_1 \cdot \hat{\mu}_2 < 0 \text{ and } \min\{|\hat{\mu}_1|/\sigma_1, |\hat{\mu}_2|/\sigma_2\} \geq \Phi(1-\alpha).$$

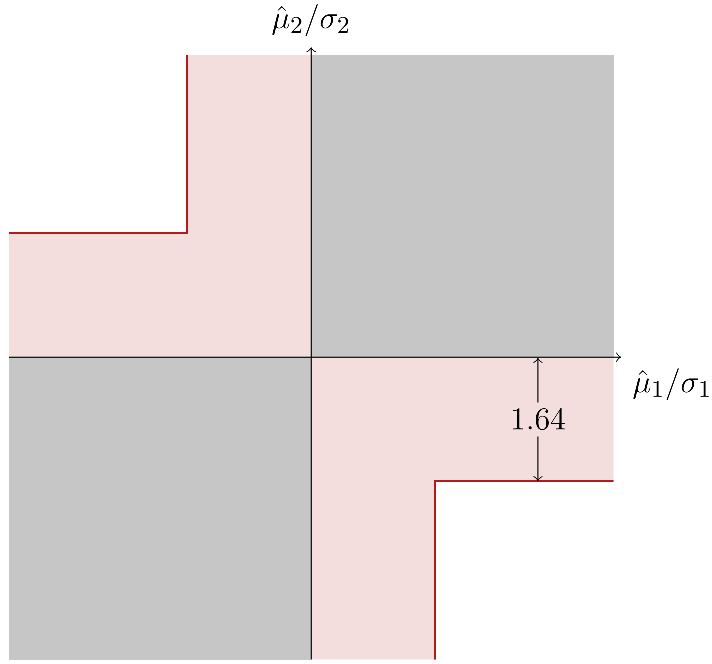


FIGURE 2.—Visualization of null hypothesis (positive and negative orthant, gray) and nonrejection region (null hypothesis plus additional shaded area, red) for the recommended test. For $\rho \geq 0$ and $\alpha = 5\%$, the test's critical value is $\Phi^{-1}(.95) \approx 1.64$.

In what follows, we call a test *valid* if it controls size. Formally, let the sample statistic $\phi(\cdot)$ be an indicator of whether the test rejects. Then validity requires that

$$\sup_{\mu_1,\mu_2,\sigma_1,\sigma_2,\rho:\mu_1\mu_2\geq 0} E(\phi(\hat{\mu}_1,\hat{\mu}_2)) \leq \alpha. \tag{2.5}$$

We call a test *nonconservatively valid* if equality holds.

THEOREM 2.1: *Suppose* (2.1) *holds. Then the test in* (2.3) *is nonconservatively valid.*

We provide an independent proof of this result in Appendix A.1. To build intuition, consider the visualization of the null hypothesis (gray) and nonrejection region (gray and red) of the test in Figure 2. Suppose $\rho = 0$ and $\sigma_1 = \sigma_2 = 1$, so $(\hat{\mu}_1, \hat{\mu}_2)$ are bivariate standard normal but centered at $(\mu_1, \mu_2)$. Suppose also that $\alpha = .05$. The claim is that, for $(\mu_1, \mu_2)$ anywhere in the gray area, the probability of $(\hat{\mu}_1, \hat{\mu}_2)$ hitting the unshaded area is at most $\alpha$ and that this bound is attained in the worst case. To see that this is so, one can apply Anderson's Lemma separately on each ascending $45^\circ$ line to show that rejection probability is maximized on the boundary of $H_0$, i.e. for $(\mu_1, \mu_2)$ on some axis. Without further loss of generality, suppose $\mu_1 = 0$, i.e. the true parameter vector is on the vertical axis. Now imagine flipping the third and fourth quadrant in Figure 2. If $\mu_1 = 0$, this does not affect the test's rejection probability. But the nonrejection region now contains all of $\{\hat{\mu}_1 \geq -1.64\}$; hence, rejection implies rejection of the one-sided null that $\mu_1 \geq 0$ and so cannot exceed $5\%$.

The last argument can be extended to $\rho \geq 0$, but it is not available for $\rho < 0$ and $c^*$ must then be calibrated. It will approach $\Phi(1-\alpha/2)$ as $\rho \to -1$, explaining why a Bonferroni adjustment based test can indeed not improve on $\Phi(1-\alpha/2)$ as critical value; in this sense, BMW's approach is tight. Finally, it is easy to verify that, for parameter values $(0, \mu_2)$ where $\mu_2 \to \infty$, the test is essentially a one-sided Wald test operated on $\hat{\mu}_1$. The test is therefore not conservative – its supremal rejection probability on the null equals $1-\alpha$, formally because the slack in the last inequality above vanishes as $|\mu_2| \to \infty$.

### 2.3. *Validity of the Heuristic Bootstrap*

Validity of the heuristic bootstrap for this testing problem is far from obvious. Indeed, recall that the test is based on a simple nonparametric bootstrap approximation, but such an approximation presumes symmetry of the distribution that is being simulated (as nicely discussed in Hansen, 2022, Section 10.16). Figure 1 therefore casts doubt not only on the Delta-method but also on inference based on the percentile bootstrap. Indeed, we can report the following result.

PROPOSITION 2.1: *Without further restrictions on $(\rho, \mu_1, \mu_2)$, the true rejection probability of the heuristic bootstrap test can take any value in $(0, 1)$, irrespective of whether $H_0$ is true or not.*

This raises the question whether the heuristic bootstrap is valid in applications where it has previously been used. Fortunately, we are able to (i) establish such validity for applications where $\rho = 0$ is known and (ii) verify that this case obtains in Kowalski (2022). The first of these observations follows from the next result.

PROPOSITION 2.2: *For $\rho = 0$, the BMW rejection region is contained in the interior of the heuristic bootstrap rejection region, which in turn is in the interior of the recommended test's rejection region.*

Given validity of the recommended test, the proposition establishes a fortiori that, with $\rho = 0$, the heuristic bootstrap is valid in the sense of controlling size. This case is not without interest as it obtains when estimates $(\hat{\mu}_1, \hat{\mu}_2)$ come from independent samples or distinct subsamples of the same (exchangeable) sample. For example, this characterizes the application to meta-analyses because independent studies presumably have independent errors. Much less obviously, $\rho = 0$ also obtains in Kowalski (2022). Formally, by tedious manipulation of some expressions in Kowalski (2022), we can show the following:

PROPOSITION 2.3: *The testing problem in Kowalski (2022) is characterized by $\rho = 0$.*

Taken together, these propositions imply that all claims Kowalski (2022) makes about the test for her setting are in fact correct; in particular, the test is both valid and less conservative than Bonferroni adjustment. However, this assessment depends on $\rho = 0$. While we prove that this holds in her setting, it will not extend to more general estimators of treatment effects across groups and over time, or in the context of mediation analysis. Furthermore, the heuristic bootstrap's appeal is limited by the fact that, if it is valid, then a simpler and yet uniformly more powerful test is available.

## 3. CAN THE TEST BE FURTHER IMPROVED?

While our recommended test improves on some other tests in the literature, its power is low around $(0, 0)$ – for example, if $\rho = 0$, power at $(0, 0)$ is easily computed to be $\alpha/10$, where $\alpha$ is nominal rejection probability. By the same token, the test is biased – close to $(0, 0)$ but outside of $H_0$, rejection probability is lower than at $(\mu, 0)$ for large $\mu$, even though the latter is in $H_0$. Can we remedy these issues? It turns out that the answer is "yes" if we allow for highly uncoventional tests that we do not seriously recommend. We then impose a condition that excludes such tests and show that, under this condition, the answer is “no” and, for most values of $\rho$, the proposed test is in a strong sense best.

Let $\rho = 0$ and $\alpha = .05$. Partition the positive real line into bins with end points

$$(\Phi^{-1}(.5), \Phi^{-1}(.55), \Phi^{-1}(.6), \Phi^{-1}(.65), \ldots, \Phi^{-1}(.9), \Phi^{-1}(.95)) \\ \approx (0, 0.13, 0.25, 0.39, 0.52, 0.67, 0.84, 1.04, 1.28, 1.64),$$

where $\Phi^{-1}(\cdot)$ is the standard normal quantile function. In words, the positive real line is divided into intervals that correspond to “.05-quantile bins” of the standard normal distribution. We can then define the following test, which is visualized in Figure 3:

$$\text{Reject } H_0 \text{ if } \hat{\mu}_1 \cdot \hat{\mu}_2 < 0 \text{ and } (|\hat{\mu}_1|/\sigma_1, |\hat{\mu}_2|/\sigma_2) \text{ are in the same bin.} \tag{3.1}$$

A similar test was considered for a related, but one-sided testing problem in Berger (1989) (and criticized in Perlman and Wu (1999)). In more closely related work, van Garderen and van Giersbergen (2022) develop the same idea for $H_0 : \mu_1 \cdot \mu_2 = 0$ and so do not encounter a question of unbiasedness, but they do show essential uniqueness of this test. Numerical computation of nonlinear rejection regions with similar intuitions was proposed (at least) in Chiburis (2008), Kaplan (2015), and Hillier et al. (2022).

We next show that the test in Eq. (3.1) is asymptotically similar (its rejection probability equals $0.05$ everywhere on the boundary of $H_0$) and unbiased ($H_0$ is a lower contour set of the power function).

PROPOSITION 3.1: *Let $\rho = 0$. The test in Eq. (3.1) has rejection probability equal to $0.05$ at every point $(\mu_1, \mu_2)$ with $\mu_1 \cdot \mu_2 = 0$, i.e., uniformly on the boundary of $H_0$. The test is furthermore unbiased. Indeed, its rejection probability strictly exceeds $0.05$ everywhere outside $H_0$ and is strictly below $0.05$ in the interior of $H_0$.*

Proposition 3.1 establishes that, in an abstract statistical sense, the test has attractive properties. However, even disregarding that its construction relies on $\alpha$ evenly dividing into

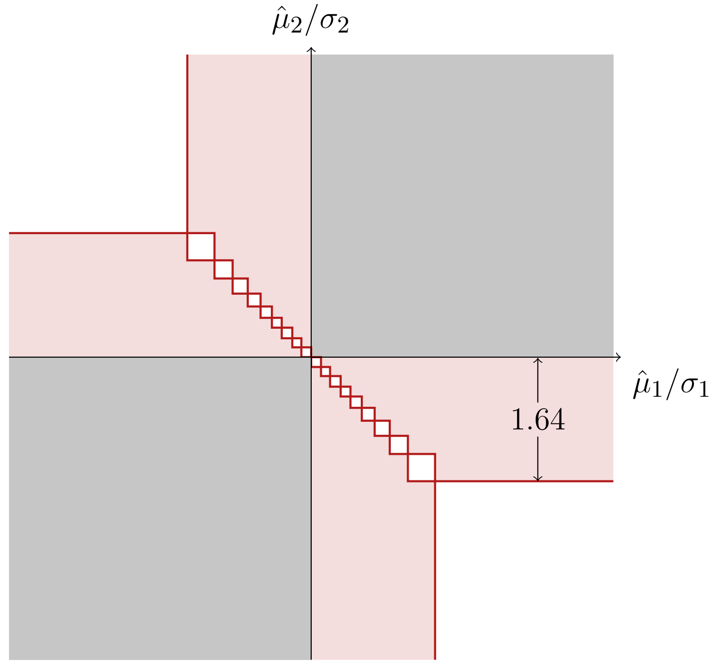


FIGURE 3.—A test with uniformly exact size control (i.e., constant rejection probability of $5\%$) on the boundary of $H_0$. Here, $H_0$ is gray (positive and negative quadrant) and the nonrejection region beyond $H_0$ is red (irregularly shaped object in second and fourth quadrants).

$1$, the test is not intuitively compelling by any stretch. For example, it rejects for $(\hat{\mu}_1, \hat{\mu}_2) = (m, -m)$ no matter how small $m$ is. This means that the data may be arbitrarily close to $H_0$ and still be deemed to reject it; by the same token, there are points in the rejection region where the p-value against the simple null hypothesis that $(\mu_1, \mu_2) = (0, 0)$ is arbitrarily close to $1$. In addition, there are fixed alternatives against which the test is not consistent, i.e. it does not asymptotically reject them with certainty.[5]

These considerations lead us to not recommend the test. However, they show that our proposed test can be most powerful only within a restricted class of tests even if basic desiderata like nontriviality (the test in fact depends on data) are imposed, and also that unbiasedness (the null hypothesis is a lower contour set of the power function) may be an unrealistic demand.

[5]These observations are related to more general points made in Andrews (2012) and Perlman and Wu (1999).

We next argue that the recommended test is best within a reasonable comparison set. To this end, consider the following definition.

DEFINITION 3.1: *A test is monotonic if, for any* $(m_1, m_2)$ *in the rejection region, one cannot leave the rejection region by increasing the absolute value of one or both of* $(m_1, m_2)$ *while leaving their signs unchanged.*

In another belatedly discovered precedent, Kaplan (2015, Section 4.2) proposes the same monotonicity condition and establishes that it brings one back to the recommended test for $\rho = 0$, a finding that we will generalize below. Intuitively, if a test is monotonic, one cannot pass from the rejection region into the nonrejection region by moving *away from* the null hypothesis. For example, in Figure 2, any "forbidden" movement would go southeast [northwest] from within the rejection region's southeastern [northwestern] component. The condition is evidently met. In contrast, the square components of the rejection region in Figure 3 violate it. We find this condition appealing.[6] That said, it excludes tests that were recommended in the aforementioned literature, and users who disagree with its heuristic appeal should certainly feel free to use such tests.

We next establish an important property of monotonic tests.

PROPOSITION 3.2: *If a monotonic test is valid in the sense of* (2.5)*, then its rejection region is contained in*

$$\boldsymbol{R}_\alpha := \{\hat{\mu}_1, \hat{\mu}_2 \;:\; \hat{\mu}_1 \cdot \hat{\mu}_2 < 0,\;\; \min\{|\hat{\mu}_1|, |\hat{\mu}_2|\} \geq \Phi^{-1}(1-\alpha)\}.$$

Whenever $c^* = \Phi^{-1}(1-\alpha)$, the recommended test's rejection region coincides with $\boldsymbol{R}_\alpha$. Hence, in this case, the recommended test is uniformly most powerful among monotonic, valid tests. In particular, this conclusion provably holds for $\rho \geq 0$ and numerically obtains

[6] This appeal may relate to the fact that, along movements described in Definition 3.1, the maximal likelihood of the data under the null hypothesis cannot increase. Note however that, while the rejection region in Figure 2 is a lower contour set of this maximized likelihood, Definition 3.1 allows for rejection regions that are not. Indeed, it is easy to see that any rejection region that is a lower contour set of the maximized likelihood corresponds to our recommended test except for possibly using a different critical value. See Kline (2011) for several other examples of counterintuitive nonmonotonicities in LR-type test statistics and, by implication, rejection regions.

for $\rho > \underline{\rho} \approx -.8$. We conclude that the test recommended here is in a strong sense optimal for most practically relevant parameter values. For the special case of strong negative correlation of estimators, we do not recommend it because, for example, a moment selection test might fare better (see, e.g., Kim, 2023).

## 4. MORE GENERAL NULL HYPOTHESES

This section sketches out a generalization of our results that is motivated by analysis in BMW. They consider a class of hypotheses where $(\mu_1, \mu_2)$ may not be constrained to the first and third quadrant but to either the positive or the negative span of two positive vectors (see Appendix C of BMW). If these vectors are known, then this is the generalization of the previous problem to null hypotheses that are the union of a polyhedral cone and its negative (with apex at the origin, though the origin itself is arbitrary).

The insight behind this generalization is that any such cone can be expressed as union of the positive and negative quadrants upon a base change. Simply, if a full dimensional cone $C$ is spanned by vectors $(b_1, b_2)$, then $\boldsymbol{A} = [b_1 \;\; b_2]^{-1}$ denotes the base change matrix under which $C$ is the positive orthant. The null hypothesis

$$H_0 : (\mu_1, \mu_2) \in C \cup -C$$

is then equivalent to

$$H_0 : (\nu_1, \nu_2) \equiv (\boldsymbol{A}\mu_1, \boldsymbol{A}\mu_2) \in \mathbb{R}^2_+ \cup \mathbb{R}^2_-.$$

The methods developed here apply, where we have that

$$(\hat{\nu}_1, \hat{\nu}_2) \sim \mathcal{N}\left(0, \boldsymbol{A}\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\boldsymbol{A}'\right).$$

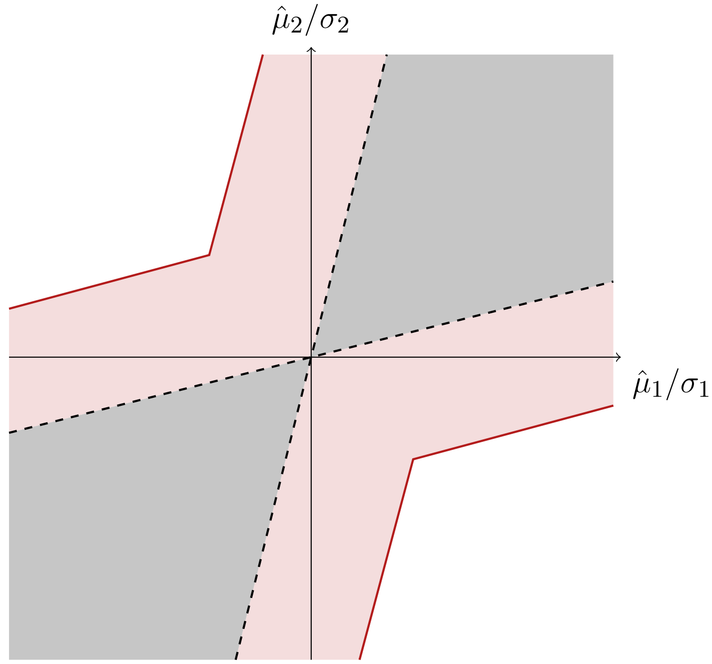


FIGURE 4.—Visualization of the recommended test for more general null hypotheses; cf. Section 4.

To illustrate this numerically, consider the cone that is spanned by $\{(2,1),(1,2)\}$. We then have that

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \implies \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} = \begin{bmatrix} 2\mu_1/3 - \mu_2/3 \\ -\mu_1/3 + 2\mu_2/3 \end{bmatrix}.$$

Therefore,

$$(\mu_1, \mu_2) \in C \cup -C \iff \nu_1 \cdot \nu_2 \geq 0,$$

and the previous analysis applies. The asymptotic version of this analysis can also easily accommodate the possibility that $(\nu_1, \nu_2)$ are approximated by consistent estimators $(\hat{\nu}_1, \hat{\nu}_2)$, as is the case in BMW. The resulting test's nonrejection region is visualized in Figure 4.

## 5. EMPIRICAL APPLICATIONS

A simple application of our results revisits the papers whose analyses motivated our work. For example, both BMW and Kowalski (2022) report Bonferroni adjusted p-values. In the latter case, it follows from our results that these p-values can be halved. For example, Kowalski (2022, pp. 448-449) carries out tests for treatment effect homogeneity and reports a p-value of $0.043$ using the BMW method and of $0.023$ using her heuristic bootstrap procedure. Our preferred test yields a p-value of $0.0215$; note how these p-values are ordered in accordance with Proposition 2.2.

The test can also be useful in the context of mediation analysis. For example, Dippel et al. (2021) examine the mediation role that labor market adjustments play in the impact of trade exposure on far-right party support. A key parameter of interest in this study is the product of $\beta_M^T$ (the effect of trade exposure on local labor markets) and $\beta_Y^M$ (the effect of the strength of the labor market on far-right party support), where each piece comes from a separate estimation model. The sign of this product gives the direction of the mediation effect. It is positive, but Dippel et al. (2021) do not remark on significance. We next verify that the indirect effect is, in fact, statistically significantly positive.

The parameters are estimated on the same data, and their estimating equations have some variables in common, so $\hat{\beta}_M^T$ and $\hat{\beta}_Y^M$ may be correlated. Practical complications in estimating this correlation include the facts that each model is estimated via 2SLS and that the data are grouped by 93 commuting zones, meaning that cluster-robust inference is warranted. We accordingly perform a cluster block bootstrap (with 10,000 replications). We sample over clusters defined by commuting zones, re-estimate the parameters within each bootstrap sample, and compute the resulting correlation between the two estimators.[7]

We focus on specifications (1) and (2) reported in Table 5 of Dippel et al. (2021). The difference between them lies in the mediating variables, with (1) defining the mediator as change in log employment and (2) defining the mediator as an aggregated labor market component derived from principal component analysis. Based on our estimation, the corre-

[7] We employ a trimmed estimator of bootstrap covariance as outlined in Hansen (2022, Section 10.14). Specifically, we set to zero the top $1\%$ of bootstrap samples ordered by the bootstrap analog of $\|(\hat{\beta}_M^T, \hat{\beta}_Y^M)\|$.

lation between estimators reported in Column 1 of Table 5 in Dippel et al. (2021) is 0.219; for the estimators in Column 2, it is 0.038. Because we can confidently assess both correlations to be greater than zero (and certainly greater than $-.75$), the one-sided critical value applies; similarly, the p-value of the test in Eq. (1.1) can be obtained as if the test were simple one-sided.

Again, Dippel et al. (2021, Table 5, Panel B) find that $\hat{\beta}_M^T \hat{\beta}_Y^M > 0$ and therefore obtain a positive point estimate for the amount of mediation, but they do not test the hypothesis of no or negative mediation and accordingly cannot (and do not) claim that their estimate is statistically significant. We can now clarify that such a claim is indeed warranted. For example, given that $\hat{\beta}_M^T \hat{\beta}_Y^M > 0$, at significance level $\alpha = 0.05$ we can compare $\min(|\hat{t}_M^T|, |\hat{t}_Y^M|)$ to $1.64$, where $\hat{t}_M^T = \hat{\beta}_M^T / \hat{\sigma}_M^T$ and $\hat{t}_Y^M = \hat{\beta}_Y^M / \hat{\sigma}_Y^M$. Using Dippel et al.'s (2021) replication package to compute these numbers, we conclude that the mediator plays a statistically significant positive role in the overall impact. Indeed, with the (bootstrapped) knowledge that the correlation between the two parameters is non-negative, the overall p-value against zero mediation can be immediately read off their Tables 4 and 5, which report p-values from (two-sided) tests of significance for $(\hat{\beta}_M^T, \hat{\beta}_Y^T)$. The recommended test's p-value is just one half of the larger of these, i.e. $0.015$ for specification (1) and $0.013$ for specification (2).

## 6. CONCLUSION

We provide simple tests for bivariate sign congruence, a problem that has recently received renewed attention in empirical research and has a long history in the literature on testing for heterogeneous treatment effects. The problem is hard because the null hypothesis is irregularly shaped (nonconvex and without interior) at the origin of parameter space. We rediscover a test from an older statistics literature that unambiguously improves on recent empirical work in the social sciences, and we clarify when some of these other proposals apply. We also argue that, for most parameter values, the test is optimal subject to some arguably plausible desiderata; also, we illustrate why some such desiderata are necessary. The results can be immediately applied to existing empirical work, e.g. by dividing p-values reported in some papers by 2. We emphasize that our result's striking simplicity owes to restricting attention to the bivariate testing problem. For $K > 2$ means with uncorrelated estimators, we recommend the likelihood ratio test in Gail and Simon (1985); for the most general case with unrestricted covariance matrix, we refer to Kim (2023).

## APPENDIX A: PROOFS

The following lemma will be used repeatedly.

LEMMA A.1: *Let Equation (2.1) hold. Consider any test such that the intersection of the rejection region with any ascending $45°$-line is a line segment centered at the descending $45°$-line; that is, for any fixed $\lambda \in \mathbb{R}$, the intersection of the rejection region with*

$$\left\{ \begin{bmatrix} -\lambda \\ \lambda \end{bmatrix} + a \begin{bmatrix} 1 \\ 1 \end{bmatrix} : a \in \mathbb{R} \right\}$$

*is either empty or equals*

$$\left\{ \begin{bmatrix} -\lambda \\ \lambda \end{bmatrix} + a \begin{bmatrix} 1 \\ 1 \end{bmatrix} : a \in [-a^*(\lambda), a^*(\lambda)] \right\}$$

*for some $a^*(\lambda) \in [0, \infty]$. Then the supremum of the test's rejection probability on $H_0$ (i.e., its size) equals its supremum on the boundary of $H_0$.*

PROOF: The null hypothesis in Eq. (1.1) can equivalently be written as

$$H_0' : \frac{\mu_1}{\sigma_1} \cdot \frac{\mu_2}{\sigma_2} \geq 0.$$

For the remainder of this proof, whenever talking about the null hypothesis we refer to $H_0'$, and to simplify expressions we set $\sigma_1 = \sigma_2 = 1$. Observing that we can reparameterize $(\mu_1, \mu_2) = (-\lambda + a, \lambda + a)$, we show below that, for any $\lambda$, rejection probability strictly decreases in $|a|$. Since no point in the interior of $H_0'$ corresponds to $a = 0$, it follows that for any interior point, there is a point on the boundary of $H_0'$ that induces strictly higher rejection probability. This establishes the lemma.

To show that, for any $\lambda$, the rejection probability strictly decreases in $|a|$, let $RR$ denote a rejection region for a test such that the assumption of the lemma holds. Then

$$RR(\lambda) \equiv RR \cap \left\{ \begin{bmatrix} -\lambda \\ \lambda \end{bmatrix} + a \begin{bmatrix} 1 \\ 1 \end{bmatrix} : a \in \mathbb{R} \right\} = \left\{ \begin{bmatrix} -\lambda \\ \lambda \end{bmatrix} + a \begin{bmatrix} 1 \\ 1 \end{bmatrix} : a \in [-a^*(\lambda), a^*(\lambda)] \right\}$$

is a (possibly unbounded) segment centered at $[-\lambda, \lambda]$ with slope equal to 1, and $a^*(\lambda) \in [0, \infty]$. Let $(X_1, X_2)$ be random variables such that

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} -\lambda \\ \lambda \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Then the claim is further equivalent to showing that for any fixed $\lambda \in \mathbb{R}$, $\Pr((X_1, X_2) \in RR) \geq \Pr((X_1, X_2) + (a, a) \in RR)$, for any $a \in \mathbb{R}$. To see that this is the case, let $Z_1 = X_1 + X_2 + 2a$ and $Z_2 = X_1 - X_2$, so that $(Z_1, Z_2)$ are uncorrelated bivariate Normal random variables, with mean, respectively, $2a$ and $-2\lambda$, and variance, respectively, $2(1+\rho)$ and $2(1-\rho)$, and $Z_1 | Z_2 \sim \mathcal{N}(2a, 2(1+\rho))$. Next, note that

$$\Pr((X_1, X_2) \in RR) = \int_{-\infty}^{+\infty} \Pr((X_1, X_2) \in RR | Z_2 = z) dF_{Z_2}(z),$$

with $F_{Z_2}$ the distribution of $Z_2$, and that

$$\Pr((X_1, X_2) \in RR | Z_2 = z) = \begin{cases} 0 & \text{if } z < \hat{c}/2 \\ \Pr(2\hat{c} - z \leq Z_1 \leq z - 2\hat{c} | Z_2 = z) & \text{if } z \geq \hat{c}/2 \end{cases}$$

By Anderson's Lemma, observing that the interval $[2\hat{c} - z, z - 2\hat{c}]$ is convex and symmetric, the above probability is maximized at $a = 0$ for any $\lambda$ and strictly decreases in $|a|$. *Q.E.D.*

### A.1. *Proof of Theorem 2.1*

Fix $\hat{c}$ and, without loss of generality (see Lemma A.1), set $\sigma_1 = \sigma_2 = 1$. The test's rejection probability is (recall notation $(\wedge, \vee)$ for (and,or))

$$\gamma(\mu_1, \mu_2) \equiv \Pr(\hat{\mu}_1 \hat{\mu}_2 < 0 \wedge |\hat{\mu}_1| > \hat{c} \wedge |\hat{\mu}_2| > \hat{c}).$$

We claim that

$$\sup\{\gamma(\mu_1, \mu_2) : \mu_1 \mu_2 \geq 0\} = \sup\{\gamma(\mu_1, \mu_2) : \mu_1 = 0, \mu_2 \geq 0\},$$

justifying the proposed test.

By symmetry of the testing problem in parameter space, we can restrict attention to $\mu_1 \geq 0$ and $\mu_2 \geq \mu_1$. Within this set, for any $\delta \geq 0$, $\gamma(\mu_1, \mu_2) \geq \gamma(\mu_1 + \delta, \mu_2 + \delta)$ by Lemma A.1. It follows that the above supremum is attained on $\{0\} \times [0, \infty)$, possibly in the limit as $\mu_2 \to \infty$.

Next, fix $\rho \geq 0$. Define

$$(X_1, X_2) \equiv (\hat{\mu}_1 - \mu_1, \hat{\mu}_2 - \mu_2) \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Also recalling that $\mu_1 = 0$, write

$$\begin{aligned}
&\Pr(\hat{\mu}_1 \hat{\mu}_2 < 0 \wedge |\hat{\mu}_1| > \hat{c} \wedge |\hat{\mu}_2| > \hat{c}) \\
&= \Pr\big([\hat{\mu}_1 < -\Phi(1-\alpha) \wedge \hat{\mu}_2 > \Phi(1-\alpha)] \vee [\hat{\mu}_1 > \Phi(1-\alpha) \wedge \hat{\mu}_2 < -\Phi(1-\alpha)]\big) \\
&= \Pr\big([X_1 < -\Phi(1-\alpha) \wedge \mu_2 + X_2 > \Phi(1-\alpha)] \vee [X_1 > \Phi(1-\alpha) \wedge \mu_2 + X_2 < -\Phi(1-\alpha)]\big) \\
&\leq \Pr\big([X_1 < -\Phi(1-\alpha) \wedge \mu_2 + X_2 > \Phi(1-\alpha)] \vee [X_1 < -\Phi(1-\alpha) \wedge \mu_2 + X_2 < -\Phi(1-\alpha)]\big) \\
&< \Pr\big(X_1 < -\Phi(1-\alpha)\big) = \alpha,
\end{aligned}$$

where the weak inequality used that $\rho \geq 0$. The supremum can be approximated by letting $\mu_2 \to \infty$.

### A.2. *Proof of Proposition 2.1*

One can easily verify that the rejection probability in question approaches $1$ as $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \to (0, 0, 1, 1, -1)$, whereas it approaches $0$ as $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \to (0, 0, 1, 1, 1)$. This is irrespective of whether $(\mu_1, \mu_2)$ approach their limits from inside or outside of $H_0$.

### A.3. *Proof of Proposition 2.2*

We again set $\sigma_1 = \sigma_2 = 1$. For the BMW test, the claim is obvious (for any $\rho$). For the heuristic bootstrap, fix $\rho = 0$ and consider any $(\hat{\mu}_1, \hat{\mu}_2)$ on the boundary of our test's rejection region. Without further loss of generality, set $\hat{\mu}_1 = \Phi(1-\alpha) > 0$. Defining

$$(X_1^*, X_2^*) = (\hat{\mu}_1^* - \hat{\mu}_1, \hat{\mu}_2^* - \hat{\mu}_2) \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

one can write

$$\begin{aligned}
\Pr(\hat{\mu}_1^* \cdot \hat{\mu}_2^* > 0) &= \Pr((\hat{\mu}_1 + X_1^*)(\hat{\mu}_2 + X_2^*) > 0) \\
&= \Pr\big([X_1^* > -\hat{\mu}_1 \wedge X_2^* > -\hat{\mu}_2] \vee [X_1^* < -\hat{\mu}_1 \wedge X_2^* < -\hat{\mu}_2]\big) \\
&> \Pr\big([X_1^* < -\hat{\mu}_1 \wedge X_2^* > -\hat{\mu}_2] \vee [X_1^* < -\hat{\mu}_1 \wedge X_2^* < -\hat{\mu}_2]\big) \\
&= \Pr\big(X_1^* < -\hat{\mu}_1\big) = 1 - \alpha.
\end{aligned}$$

It follows that the entire boundary of the recommended test's rejection region is in the interior of the heuristic bootstrap test's *nonrejection* region.

To see that the BMW rejection region is interior to the heuristic bootstrap one, consider maximizing the heuristic bootstrap p-value of $(\hat{\mu}_1, \hat{\mu}_2)$ on the BMW rejection region. It is easy to see that this problem is solved by $(\Phi(1-\alpha/2), -\Phi(1-\alpha/2))$. But at that point, the heuristic bootstrap p-value equals $2 \cdot \alpha/2 - (\alpha/2)^2 < \alpha$, so that it is interior to the heuristic bootstrap rejection region.

## A.4. *Proof of Proposition 2.3*

In this proof only, all notation is exactly as in Kowalski (2022). Substitute

$$\begin{aligned}
E(Y_T|A) &= E(Y|D=1,Z=0)\\
E(Y_U|N) &= E(Y|D=0,Z=1)\\
E(Y_T|C) &= \tfrac{p_I}{p_I-p_C}E(Y|D=1,Z=1)-\tfrac{p_C}{p_I-p_C}E(Y|D=1,Z=0)\\
E(Y_U|C) &= \tfrac{1-p_C}{p_I-p_C}E(Y|D=0,Z=0)-\tfrac{1-p_I}{p_I-p_C}E(Y|D=0,Z=1)
\end{aligned}$$

into

$$\tau \equiv \big(E(Y_U|C)-E(Y_U|N)\big)\cdot\big(E(Y_T|A)-E(Y_T|C)\big)$$

to find

$$\begin{aligned}
\tau &= \big(\tfrac{1-p_C}{p_I-p_C}E(Y|D=0,Z=0)-\tfrac{1-p_I}{p_I-p_C}E(Y|D=0,Z=1)-E(Y|D=0,Z=1)\big)\\
&\quad\cdot\big(E(Y|D=1,Z=0)-\tfrac{p_I}{p_I-p_C}E(Y|D=1,Z=1)+\tfrac{p_C}{p_I-p_C}E(Y|D=1,Z=0)\big)\\
&= \tfrac{1-p_C}{p_I-p_C}\big(E(Y|D=0,Z=0)-E(Y|D=0,Z=1)\big)\\
&\quad\cdot\tfrac{p_I}{p_I-p_C}\big(E(Y|D=1,Z=0)-E(Y|D=1,Z=1)\big)\\
&= \tfrac{p_I(1-p_C)}{(p_I-p_C)^2}\big(E(Y|D=0,Z=0)-E(Y|D=0,Z=1)\big)\\
&\quad\cdot\big(E(Y|D=1,Z=0)-E(Y|D=1,Z=1)\big),
\end{aligned}$$

so that it suffices to sign

$$[(E(Y|D=0,Z=0)-E(Y|D=0,Z=1)]\cdot[(E(Y|D=1,Z=0)-E(Y|D=1,Z=1)],$$

but these expectations condition on mutually exclusive events and their sample analogs are computed from mutually exclusive subsamples.

### A.5. *Proof of Proposition 3.1*

Again set $\sigma_1 = \sigma_2 = 1$. To see uniform validity, w.l.o.g. restrict attention to parameter value $(\mu, 0)$. Let $\phi(\cdot)$ be the indicator function for the test in Eq. (3.1). Then for any $m_1 \neq 0$, it is easily seen that $\Pr(\phi(\hat{\mu}_1, \hat{\mu}_2) = 1 \mid \hat{\mu}_1 = m_1) = .05$. (In particular, consider that $\hat{\mu}_2$ is standard normal and that conditionally on any $\hat{\mu}_1 \neq 0$, the probability of $(\hat{\mu}_1, \hat{\mu}_2)$ falling in the rejection region is the probability of $\hat{\mu}_2$ hitting some .05-quantile bin of the standard normal distribution.)

To see unbiasedness, observe that Lemma A.1 applies; hence, for any fixed value of $(\mu_2 - \mu_1)$, probability of rejection strictly decreases in $|\mu_1 + \mu_2|$. This establishes the remainder of the proposition's claims.

### A.6. *Proof of Proposition 3.2*

Suppose by contradiction that the rejection region contains a point $(m_1, m_2)$ with $m_1 > 0$ and $m_2 > -\Phi^{-1}(1 - \alpha)$. Then by monotonicity, it also contains $(m_1 + \lambda, m_2)$ for any $\lambda \geq 0$. But this implies that, along any sequence of parameter values $\{(0, \mu_{2,n}) : \mu_{2,n} \to \infty\}$, rejection probability converges to (at least) $\Phi(-m_2) > \alpha$. The argument for other possible values of $(m_1, m_2)$ is similar.

## REFERENCES


ANDREWS, DONALD (2012): "Similar-on-the-Boundary Tests for Moment Inequalities Exist, But Have Poor Power," Cowles Foundation Discussion Papers 1815R, Cowles Foundation for Research in Economics, Yale University. [12]

BERGER, ROGER L. (1989): "Uniformly More Powerful Tests for Hypotheses Concerning Linear Inequalities and Normal Means," *Journal of the American Statistical Association*, 84, 192–199. [11]

BRINCH, CHRISTIAN N., MAGNE MOGSTAD, AND MATTHEW WISWALL (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039. [2, 3, 6, 9, 10, 14, 15, 16, 21]

CHIBURIS, RICHARD (2008): "Approximately Most Powerful Tests for Moment Inequalities," Working paper, Princeton University. [11]

CRAIG, CECIL C. (1936): "On the Frequency Function of $xy$," *The Annals of Mathematical Statistics*, 7, 1 – 15. [5]

DATTA, SOMNATH (1995): "On a modified bootstrap for certain asymptotically nonnormal statistics," *Statistics & Probability Letters*, 24, 91–98. [5]

DIPPEL, CHRISTIAN, ROBERT GOLD, STEPHAN HEBLICH, AND RODRIGO PINTO (2021): "The Effect of Trade on Workers and Voters," *The Economic Journal*, 132, 199–217. [1, 16, 17]

EFRON, B. (1979): "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26. [6]

GAIL, M. AND R. SIMON (1985): "Testing for Qualitative Interactions between Treatment Effects and Patient Subsets," *Biometrics*, 41, 361–372. [7, 18]

HANSEN, BRUCE (2022): *Econometrics*, Princeton University Press. [9, 16]

HILLIER, GRANT, KEES JAN VAN GARDEREN, AND NOUD VAN GIERSBERGEN (2022): "Improved tests for mediation," Tech. rep., cemmap working paper CWP01/22. [11]

KAPLAN, DAVID M. (2015): "Bayesian and frequentist tests of sign equality and other nonlinear inequalities," Working Paper 1516, University of Missouri. [3, 11, 13]

KAPLAN, DAVID M. AND LONGHAO ZHUO (2021): "Frequentist properties of Bayesian inequality tests," *Journal of Econometrics*, 221, 312–336. [3]

KIM, DEBORAH (2023): "Testing Sign Agreement," available at https://www.dropbox.com/scl/fi/e8b9o1bt9ful8s21zh5yi/JMP_DKIM.pdf?rlkey=8rsfnc2572e4gnopflj1y1mlc&dl=0. [2, 3, 4, 14, 18]

KLINE, BRENDAN (2011): "The Bayesian and frequentist approaches to testing a one-sided hypothesis about a multivariate mean," *Journal of Statistical Planning and Inference*, 141, 3131–3141. [13]

KOWALSKI, AMANDA E (2022): "Behaviour within a Clinical Trial and Implications for Mammography Guidelines," *The Review of Economic Studies*, 90, 432–462. [1, 2, 3, 4, 6, 10, 16, 22]

LIU, YIQI AND FRANCESCA MOLINARI (2024): "Inference for an Algorithmic Fairness-Accuracy Frontier," available at https://arxiv.org/abs/2402.08879. [2]

MACKINNON, DAVID, CHONDRA LOCKWOOD, JEANNE HOFFMAN, STEPHEN WEST, AND VIRGIL SHEETS (2002): "MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects," *Psychological methods*, 7, 83–104. [2]

MAMMEN, ENNO (1992): *When does bootstrap work : Asymptotic results and simulations*, Springer-Verlag. [6]

MONTOYA, AMANDA KAY AND ANDREW F. HAYES (2017): "Two-condition within-participant statistical mediation analysis: A path-analytic framework." *Psychological methods*, 22 1, 6–27. [4]

PERLMAN, MICHAEL D. AND LANG WU (1999): "The Emperor's New Tests," *Statistical Science*, 14, 355–369. [11, 12]

RUSSEK-COHEN, ESTELLE AND RICHARD M. SIMON (1993): "Qualitative Interactions in Multifactor Studies," *Biometrics*, 49, 467–477. [3, 4, 7]

SLOUGH, TARA AND SCOTT A. TYSON (2024): "Sign-Congruence, External Validity, and Replication," *Political Analysis*, in press. [2, 3]

SOBEL, MICHAEL E. (1982): "Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models," *Sociological Methodology*, 13, 290. [5]

VAN GARDEREN, KEES JAN AND NOUD VAN GIERSBERGEN (2022): "A Nearly Similar Powerful Test for Mediation," ArXiv 2012.11342. [11]