Multiple Testing of Stochastic Monotonicity

Qian Wu* David M. Kaplan†

January 16, 2025

Abstract

We develop multiple testing methodology for stochastic monotonicity with an ordinal, discrete, or discretized covariate. Rather than testing a single global null hypothesis, we use multiple testing to evaluate specifically where the conditional distributions are consistent with the outcome variable stochastically increasing in the covariate. By inverting our multiple testing procedure that controls the familywise error rate, we construct "inner" and "outer" confidence sets for the true set of points consistent with stochastic increasingness. Simulations show reasonable finite-sample properties. Empirically, we apply our methodology to the relationship between mental health and education, and between earnings and education. Practically, we provide code implementing our multiple testing procedure and replicating our empirical results.

JEL classification: C25, I10

Keywords: Confidence set; Familywise error rate; Mental health; Multiple testing procedure

*School of Statistics, Southwestern University of Finance and Economics, wuqj@swufe.edu.cn

†Corresponding author; Department of Economics, University of Missouri, kaplandm@missouri.edu

1 Introduction

A fundamental empirical question is whether outcome variable Y is "increasing" in covariate X, which can be characterized in terms of stochastic monotonicity. For example, is mental health increasing in education? Variable Y is "stochastically increasing" in X if for any two X values, the conditional distribution of Y at the higher X value first-order stochastically dominates the conditional distribution of Y at the lower X value. Economically, assuming Y is scaled so higher values are better, this conditional stochastic dominance means that the conditional distribution of Y is "better" at higher X values, in the sense of higher expected utility. However, because this is a strong property, it can be statistically difficult to find strong evidence in its favor, and conversely it may nearly hold but be rejected for an economically small violation.

We contribute to the stochastic monotonicity inference literature by proposing and justifying new methodology that focuses on multiple testing, as well as allowing either continuous or ordinal/discrete outcomes. Other stochastic monotonicity tests (see below) assume a continuous outcome and only test a single null hypothesis of global stochastic monotonicity. Instead of testing the single null hypothesis that Y is stochastically increasing in X, we consider multiple testing of the conditional CDF inequalities that jointly comprise stochastic increasingness. Specifically, each inequality checks whether the conditional CDF of Y is decreasing as X increases to the next-highest value, over all values in the support of Y and any finite number of X values. Despite this being an infinite number of points when Y is continuous, our methods still control the familywise error rate.

Complementing global testing of a single null hypothesis, multiple testing addresses the aforementioned statistical concerns in the following two ways. First, if stochastic increasingness is rejected, then multiple testing shows more precisely whether many of the constituent conditional CDF inequalities are rejected, or only a few, or even just one. Second, if stochastic increasingness is not rejected, then multiple testing can be used on the reversed inequalities (conditional CDF increasing with X) to see if they can be rejected in favor of the original inequalities (decreasing with X). This gathers stronger statistical evidence in favor of stochastic increasingness than simply failing to reject the original null hypothesis, because type II error rates (false non-rejection) are not controlled. For example, due to small sample size (high statistical uncertainty), stochastic increasingness may fail to be rejected even if the estimated conditional distributions are stochastically *decreasing*.

Although the economic interpretation is the same either way, our statistical approach depends on the type of outcome variable. For ordinal or discrete outcomes, there are a finite number of inequalities, so we use a maximum t-statistic approach. This allows us to exactly control the asymptotic familywise error rate under the least favorable null. For continuous outcomes, we achieve finite-sample control of the familywise error rate by building on the two-sample multiple testing procedure of Goldman and Kaplan (2018). They use the probability integral transform and joint distribution of order statistics to achieve finite-sample control of familywise error rate. Their method applies directly to our setting if X is binary. Otherwise, we use a Bonferroni correction to maintain the finite-sample control of familywise error rate. Although the Bonferroni correction errs on the conservative side, there are two reasons this may not have a large quantitative effect on power. First, the effect is smaller with a smaller number of X values, which applies to the education variable in our empirical illustration, for example. Second, the effect is smaller if the dependence across X is more negative. Note Bonferroni is not conservative in the extreme case of “perfect negative dependence” where a false rejection at one X implies no false rejections at any other X. In our case, if the estimated conditional CDF at $X = 2$ is higher than the true one, then it makes us more likely to reject the comparisons of $X = 1$ with $X = 2$, but less likely to reject the comparisons of $X = 2$ with $X = 3$; that is, negative dependence.

Additionally, our multiple testing procedure can be inverted into “inner” and “outer” confidence sets. The inner confidence set contains all points at which the reversed inequalities have been rejected, i.e., where there is strong evidence in favor of the original conditional CDF inequality consistent with stochastic increasingness. The outer confidence set contains all points at which the original inequalities have not been rejected. Such confidence sets are similar to those of Kaplan (2024) and to equation (1) of Armstrong and Shen (2015). The inner confidence set can be seen as a conservative estimator of the true set of inequalities consistent with stochastic increasingness, in that the inner set is contained within the true set with high probability. Conversely, the outer set contains the true set with high probability.

Our methodology is illustrated through two empirical applications. Our first application has an ordinal outcome variable, looking at the relationship between depression and education. Ordinal variables are common in health, including measures of mental health. We find especially strong evidence of mental health improving with education at the margins of attaining a high-school or college (bachelor's) degree; that is, we reject that depression (which is bad) is stochastically increasing in favor of the conclusion that depression is stochastically decreasing in education at those margins. Our second application has a continuous outcome variable, looking at the relationship between individual earnings and education. We do not reject any inequality consistent with earnings stochastically increasing with education, and we find strong evidence of stochastic increasingness across a wide range of earnings at every education level.

Besides intrinsic interest in settings like the education gradient in health or intergenerational mobility (comparing child's outcome Y with their parent's outcome X), stochastic monotonicity also appears in certain identifying assumptions or as a testable implication thereof. For example, stochastic monotonicity is implied by the combination of (semi-)monotone treatment response and exogenous treatment selection, as discussed by Manski (1997, §3.4). As another example, Small, Tan, Ramsahai, Lorch, and Brookhart (2017) identify a weighted average treatment effect when the treatment is stochastically increasing in the instrument. Although testing is trivial in the binary–binary setting they focus on, more generally our methodology can help assess their identifying assumption of stochastic monotonicity.

Literature Our methodology contributes to the literature on stochastic monotonicity and multiple testing. Inference on stochastic monotonicity has focused on continuous Y and testing the single hypothesis of global stochastic monotonicity; for example, see Lee, Linton, and Whang (2009), Seo (2018), and Chetverikov, Liao, and Chernozhukov (2021). Our reversing the direction of null hypothesis inequalities to find stronger evidence in favor of stochastic increasingness is inspired by Davidson and Duclos (2013), who make a related argument for testing the null of non-dominance against the alternative of stochastic dominance, which can be seen as a special case of stochastic monotonicity with binary X. Although we emphasize multiple testing, testing the set of stochastic monotonicity inequalities jointly would fit in the (moment) inequality literature; we essentially follow the least favorable approach with a max-t statistic as in Section 4.1.1 of Canay and Shaikh (2017), whose survey includes additional references. Our proof strategies are also similar to those of Zhao and Kaplan (2024), who instead consider multiple testing of a function's value across a finite set of points. Kaplan and Zhao (2023) also use multiple testing with ordinal outcome variables, but they only compare two (unconditional) distributions and focus not on ordinal categories but on a latent variable's quantiles. Surveys of the multiple testing literature can be found in Lehmann and Romano (2005b) and Romano, Shaikh, and Wolf (2010).

Paper structure Section 2 describes the setting with an ordinal or discrete outcome, and our methodology and its properties. Section 3 describes our contributions with a continuous outcome. Section 4 applies our methodology empirically. Section 5 presents simulation results. Appendix A collects proofs.

Notation and abbreviations Generally, scalars, vectors, and matrices are respectively typeset like X, $\boldsymbol{X}$, and $\underline{\boldsymbol{X}}$. The indicator function is $\mathbb{1}\{\cdot\}$, with $\mathbb{1}\{A\} = 1$ if event A occurs and $\mathbb{1}\{A\} = 0$ if not, and "$\iff$" means if and only if. Acronyms used include those for confidence set (CS), continuous mapping theorem (CMT), cumulative distribution function (CDF), familywise error rate (FWER), and multiple testing procedure (MTP).

2 Results for ordinal and discrete outcomes

In this section, we describe the setting, assumptions, methodology, and asymptotic properties when the outcome variable is ordinal or discrete.

2.1 Setting

Consider random variables Y and X. The outcome Y is discrete or ordinal, with categories labeled $Y \in \{1, 2, \ldots, J\}$, for finite J. For example, if the ordinal categories are “disagree,” “neutral,” and “agree,” then they are respectively coded as $Y = 1$, $Y = 2$, and $Y = 3$; or if the possible discrete values are 0, 0.5, 1, and 1.5, then these are respectively coded as $Y = 1$, $Y = 2$, $Y = 3$, and $Y = 4$. The covariate may also be discrete or ordinal, with categories labeled $X \in \{1, 2, \ldots, K\}$, for finite K. Alternatively, the covariate may be a discretization of a continuous variable; for example, a continuous covariate with support $[0, 100]$ may be discretized with values in $[0, 10]$ coded as $X = 1$, $(10, 20]$ coded as $X = 2$, etc. In sum, the supports $\mathcal{Y}$ and $\mathcal{X}$ are

$$Y \in \mathcal{Y} \equiv \{1, \ldots, J\}, \quad X \in \mathcal{X} \equiv \{1, \ldots, K\}. \tag{1}$$

The population and empirical (estimated) conditional CDF values are respectively

$$\begin{aligned} F_x(y) &\equiv \mathrm{P}(Y \le y \mid X = x), \\ \hat{F}_x(y) &\equiv \frac{\sum_{i=1}^{n} \mathbb{1}\{Y_i \le y\} \mathbb{1}\{X_i = x\}}{n_x}, \quad n_x \equiv \sum_{i=1}^{n} \mathbb{1}\{X_i = x\}, \end{aligned} \tag{2}$$

given iid observations over $i = 1, \ldots, n$. That is, $F_x(y)$ is the proportion of the $X = x$ subpopulation with $Y \le y$, and $\hat{F}_x(y)$ is the proportion of the $X_i = x$ subsample with $Y_i \le y$.

We assume iid sampling and condition on the size of each subsample n_x, which is equivalent to conditioning on all the observed X_i values (like in classical linear regression results). This is also equivalent to repeated sampling of n_x iid draws of Y_i from the corresponding conditional distribution independently for each $x = 1, \ldots, K$, which is how we formalize the setting in Assumption A1.

Assumption A1. Using the notation in (1) and (2), $n_1/n_x \to \gamma_x \in (0, \infty)$ for each $x \in \mathcal{X}$, and the sample consists of n_x observations with $X_i = x$ for each $x \in \mathcal{X}$, with the corresponding Y_i values sampled iid from the corresponding population conditional distribution of $Y \mid X = x$, and all Y_i are mutually independent.

2.2 Stochastic monotonicity inequalities

Outcome Y stochastically increasing in X means that for any $x_2 > x_1$, the conditional distribution of $Y \mid X = x_2$ first-order stochastically dominates that of $Y \mid X = x_1$. In the notation of (2), this means

$$F_{x_2}(y) \leq F_{x_1}(y) \text{ for all } (x_1, x_2, y) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \text{ with } x_2 > x_1. \tag{3}$$

As with stochastic dominance testing, despite being economically different, strict and weak inequalities are statistically indistinguishable, so we do not emphasize the difference.

We test a subset of the inequalities in (3). We restrict attention to $y \in \{1, \ldots, J-1\}$ because $F_{x_2}(J) = F_{x_1}(J) = 1$ for all x_1, x_2. In principle, we could test all $\{(x_1, x_2) \in \mathcal{X} \times \mathcal{X} : x_1 < x_2\}$, which is $K(K-1)/2$ different values of (x_1, x_2). However, without claiming optimality against all alternatives, we restrict attention to the $K-1$ pairs satisfying $x_2 = x_1 + 1$ because reducing the number of comparisons from $(J-1)K(K-1)/2$ to $(J-1)(K-1)$ allows a lower critical value that improves power, while still testing a subset of inequalities jointly equivalent to (3). That is, (3) holds if and only if

$$\theta_{x,y} \leq 0 \text{ for all } (x, y) \in \{1, \ldots, K-1\} \times \{1, \ldots, J-1\}, \quad \theta_{x,y} \equiv F_{x+1}(y) - F_x(y). \tag{4}$$

Additionally, testing (4) yields results that are easier to interpret and communicate.

For ordinal outcomes, although first-order stochastic dominance generally suggests one distribution is "better" than another, the relationship is more complex if utility depends not on the observed ordinal value but rather on a latent continuous variable. We focus here on testing the conditional ordinal distributions; for interpretation with respect to a latent variable, see Kaplan and Zhao (2023).

2.3 Multiple testing procedure

Based on (4), we consider multiple testing of the following family of hypotheses:

$$H_{0x,y} \colon \theta_{x,y} \le 0 \text{ for } (x,y) \in \{1,\ldots,K-1\} \times \{1,\ldots,J-1\}. \tag{5}$$

The number of hypotheses is $(J-1)(K-1)$. A multiple testing procedure makes a binary decision (reject or not) for each hypothesis, hence $2^{(J-1)(K-1)}$ possible results.

Our multiple testing procedure (MTP) controls the asymptotic familywise error rate (FWER). Although there are other measures of overall false positive rate for multiple testing, like the k-FWER and false discovery proportion (Lehmann and Romano, 2005a) and the false discovery rate (Benjamini and Hochberg, 1995), we use FWER because it has a clear interpretation and allows us to invert our MTP into confidence sets. FWER is defined as (Lehmann and Romano, 2005b, §9.1)

$$\text{FWER} \equiv \text{P}(\text{reject any true } H_{0x,y}). \tag{6}$$

Conversely, $1 - \text{FWER}$ is the probability of having zero false rejections (zero type I errors). Our MTP has "strong control" because it controls asymptotic FWER regardless of which $H_{0x,y}$ are true or false (Lehmann and Romano, 2005b, §9.1).

Our MTP uses standard t-statistics but with a higher critical value that accounts for multiple testing. For any real scalar d, define

$$\hat{t}_{x,y}(d) \equiv \frac{\hat{F}_{x+1}(y) - \hat{F}_x(y) - d}{\hat{s}_{x,y}}, \tag{7}$$

where $\hat{s}_{x,y}$ is defined in (22) as an estimator of the standard deviation of $\hat{F}_{x+1}(y) - \hat{F}_x(y)$. As usual, in practice we compute the t-statistic centered at our hypothesized value $d = 0$, and asymptotic properties can be bounded by the behavior of the t-statistic centered at the true $d = \theta_{x,y} \equiv F_{x+1}(y) - F_x(y)$. As an important ingredient of our FWER derivations, Lemma 1 establishes the asymptotic normal distribution of random vector $\hat{\boldsymbol{t}}$ that contains all the t-statistics centered at the true $\theta_{x,y}$:

$$\hat{\boldsymbol{t}} \equiv (\hat{\boldsymbol{t}}_1, \hat{\boldsymbol{t}}_2, \ldots, \hat{\boldsymbol{t}}_{J-1})', \quad \hat{\boldsymbol{t}}_j \equiv \big(\hat{t}_{K-1,j}(\theta_{K-1,j}), \hat{t}_{K-2,j}(\theta_{K-2,j}), \ldots, \hat{t}_{1,j}(\theta_{1,j})\big). \tag{8}$$

Lemma 1. *Under Assumption A1,*

$$\hat{\boldsymbol{t}} \xrightarrow{d} \boldsymbol{t} \sim \mathrm{N}(\boldsymbol{0}, \underline{\boldsymbol{\Sigma}}), \quad \underline{\boldsymbol{\Sigma}} = \underline{\boldsymbol{S}}'\underline{\boldsymbol{W}}\underline{\boldsymbol{S}},$$

where $\underline{\boldsymbol{W}}$ *and* $\underline{\boldsymbol{S}}$ *are non-random matrices defined in* (20) *and* (24).

We provide some brief intuition for the critical value whose validity is formally established in the proof of Theorem 2. Intuitively, the least favorable configuration here is when all $\theta_{x,y} = 0$: every null $H_{0x,y}$ is true (and thus can contribute to FWER), but just barely, so there is the highest probability of some estimated $\hat{\theta}_{x,y} > 0$ large enough to reject $H_{0x,y}$. In that case, $\hat{t}_{x,y}(0) = \hat{t}_{x,y}(\theta_{x,y})$, and we make a familywise error whenever the maximum t-statistic $\max_{x,y} \hat{t}_{x,y}(0)$ exceeds our critical value. Thus, using the asymptotic approximation, if we want FWER $= \alpha$ in this least favorable case, then the critical value should be the $(1-\alpha)$-quantile of the distribution of the maximum of $\boldsymbol{t}$:

$$c_\alpha \equiv (1-\alpha)\text{-quantile of } \max_{x,y} t_{x,y}, \tag{9}$$

where the $t_{x,y}$ are the elements of $\boldsymbol{t}$ and $\boldsymbol{t} \sim \mathrm{N}(\boldsymbol{0}, \underline{\boldsymbol{\Sigma}})$ by Lemma 1. In practice, the unknown $\underline{\boldsymbol{\Sigma}}$ can be replaced by consistent estimator $\hat{\underline{\boldsymbol{\Sigma}}}$:

$$\hat{c}_\alpha \equiv (1-\alpha)\text{-quantile of } \max_{x,y} \tilde{t}_{x,y}, \quad \tilde{\boldsymbol{t}} \sim \mathrm{N}(\boldsymbol{0}, \hat{\underline{\boldsymbol{\Sigma}}}). \tag{10}$$

Although an analytic formula is intractable, (10) can be simulated as in our code. Alternatively, the distribution of $\boldsymbol{t}$ could be approximated by bootstrap.

Method 1. *First, compute t-statistics* $\hat{t}_{x,y}(0)$ *as in* (7). *Second, given desired FWER level*

α, simulate the critical value $\hat{c}_\alpha$ in (10) *following the steps in Section 2.5. Third, for each (x,y), to test all $H_{0x,y}\colon \theta_{x,y} \le 0$ as in* (5)*, reject $H_{0x,y}\colon \theta_{x,y} \le 0$ if $\hat{t}_{x,y}(0) > \hat{c}_\alpha$. Alternatively, to test the reversed family of hypotheses $H^{\geq}_{0x,y}\colon \theta_{x,y} \ge 0$, reject $H_{0x,y}$ if $\hat{t}_{x,y}(0) < -\hat{c}_\alpha$.*

Method 1 includes the reversed $H^{\geq}_{0x,y}$ because their rejection provides stronger evidence of $\theta_{x,y} < 0$ than does non-rejection of $H_{0x,y}$. For example, even if $\hat{\theta}_{x,y} = 1.7$ (our best guess is a positive $\theta_{x,y}$), if there is a small sample size or otherwise high uncertainty, we may still not reject $H_{0x,y}\colon \theta_{x,y} \le 0$. In contrast, to reject $H^{\geq}_{0x,y}$, not only must we have $\hat{\theta}_{x,y} < 0$, but it must be significantly less than zero (compared to our uncertainty) for the test to control the false positive rate. In sum, non-rejection of $H_{0x,y}$ suggests the data are consistent with the hypothesis that Y is stochastically increasing in X, but rejection of $H^{\geq}_{0x,y}$ provides stronger evidence in favor of the inequalities that comprise stochastic increasingness of Y in X.

Theorem 2 theoretically justifies our MTP.

Theorem 2. *Under Assumption A1, Method 1 has strong control of asymptotic FWER at level α.*

2.4 Confidence sets

The MTP in Method 1 can be inverted into confidence sets for the true set

$$\mathcal{T} \equiv \{(x,y) : \theta_{x,y} \le 0\}. \tag{11}$$

The goal is, with high asymptotic probability, for inner confidence set $\widehat{\mathcal{CS}}_{inner}$ to be contained within the true set $\mathcal{T}$, and for outer confidence set $\widehat{\mathcal{CS}}_{outer}$ to contain $\mathcal{T}$. That is, for confidence level $1-\alpha$,

$$\begin{aligned} &\lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}_{inner} \subseteq \mathcal{T}) \ge 1-\alpha, \\ &\lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}_{outer} \supseteq \mathcal{T}) \ge 1-\alpha. \end{aligned} \tag{12}$$

Intuitively, $\widehat{\mathcal{CS}}_{inner}$ provides a conservative "estimate" of the true set $\mathcal{T}$, while $\widehat{\mathcal{CS}}_{outer}$ gives a larger "estimate" describing how large the true set might be. Corresponding to the earlier discussion of stronger and weaker evidence, the inner confidence set collects points for which the reversed $H^{\geq}_{0x,y}$ is rejected (strong evidence), whereas the outer confidence set collects points for which the original $H_{0x,y}$ is not rejected (weak evidence).

Theorem 3 formally establishes the property in (12) for our confidence sets described in Method 2.

Method 2. *First, run Method 1. The inner confidence set $\widehat{\mathcal{CS}}_{inner}$ collects all pairs of (x, y) for which the reversed null $H^{\geq}_{0x,y}\colon \theta_{x,y} \geq 0$ is rejected. The outer confidence set $\widehat{\mathcal{CS}}_{outer}$ collects all pairs of (x, y) for which the original null $H_{0x,y}\colon \theta_{x,y} \leq 0$ is not rejected.*

Theorem 3. *Under Assumption A1, Method 2 satisfies* (12).

If instead of $\mathcal{T}$ in (11) we are interested in its complement $\mathcal{T}^{\complement} \equiv \{(x, y) : \theta_{x,y} > 0\}$, then the outer CS is the complement of the inner CS for $\mathcal{T}$, and the inner CS is the complement of the outer CS for $\mathcal{T}$. This follows because the event $\widehat{\mathcal{CS}}_{inner} \subseteq \mathcal{T}$ is equivalent to the inner CS complement containing $\mathcal{T}^{\complement}$, and the event $\widehat{\mathcal{CS}}_{outer} \supseteq \mathcal{T}$ is equivalent to the outer CS complement being contained within $\mathcal{T}^{\complement}$. Thus, using the notation from (12),

$$\lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}^{\complement}_{inner} \supseteq \mathcal{T}^{\complement}) = \lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}_{inner} \subseteq \mathcal{T}) \geq 1 - \alpha,$$
$$\lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}^{\complement}_{outer} \subseteq \mathcal{T}^{\complement}) = \lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}_{outer} \supseteq \mathcal{T}) \geq 1 - \alpha.$$

2.5 Critical value simulation

We compute the critical value $\hat{c}_\alpha$ in (10) by simulation. Let

$$\hat{\underline{\boldsymbol{\Sigma}}} = \hat{\underline{\boldsymbol{S}}}'\hat{\underline{\boldsymbol{W}}}\hat{\underline{\boldsymbol{S}}}, \tag{13}$$

where $\hat{\underline{\boldsymbol{W}}}$ and $\hat{\underline{\boldsymbol{S}}}$ are the sample analogs of $\underline{\boldsymbol{W}}$ and $\underline{\boldsymbol{S}}$ from (20) and (24). The consistency result in (23) implies $\hat{\underline{\boldsymbol{S}}} \xrightarrow{p} \underline{\boldsymbol{S}}$, and by the weak law of large numbers $\hat{\underline{\boldsymbol{W}}} \xrightarrow{p} \underline{\boldsymbol{W}}$, so applying the continuous mapping theorem yields $\hat{\underline{\boldsymbol{\Sigma}}} \xrightarrow{p} \underline{\boldsymbol{\Sigma}}$.

Given $\hat{\underline{\boldsymbol{\Sigma}}}$, the simulation proceeds as follows, as implemented in our code. First, we randomly draw a vector from $\mathrm{N}(\boldsymbol{0}, \hat{\underline{\boldsymbol{\Sigma}}})$. Second, we take the maximum of this vector. Third, we repeat this process N times, collecting the N maxima. Fourth, we take the $(1-\alpha)$-quantile of these N maximum values; this is $\hat{c}_\alpha$.

With large enough N, we can achieve arbitrarily small simulation error. Larger N improves accuracy but increases computation time. Given our simulation results (Section 5), we suggest $N = 1000$ as a default. In our empirical application, we use $N = 100{,}000$ because computation time is still only a few seconds.

3 Results for continuous outcomes

In this section, we describe our contributions with continuous Y.

3.1 Setting and inequalities

The setting, notation, and Assumption A1 are the same as in Section 2.1, but now Y is continuous. To contrast with $\mathcal{Y}$ from Section 2.1, here we write $\mathbb{R}$ as the support of Y, with the understanding that the true support may be a subset of $\mathbb{R}$.

Analogous to (5), here we test

$$H_{0x,y}\colon \theta_{x,y} \le 0 \text{ for } (x,y) \in \{1,\dots,K-1\} \times \mathbb{R}, \quad \theta_{x,y} \equiv F_{x+1}(y) - F_x(y). \tag{14}$$

This is an infinite number of hypotheses, corresponding to an infinite number of rejection decisions. However, they are highly dependent across y values, so for a given x, usually the rejected $H_{0x,y}$ correspond to only a few intervals of y values. Computationally, of course, we do not compute an infinite number of test statistics, but rather compute values corresponding to each observation and interpolate in a precise way.

3.2 Methodology

Our methodology builds on that of Goldman and Kaplan (2018). Their Method 5 includes a multiple testing procedure that directly applies to our setting when X is binary, so $K = 1$.

That is, given $\theta_y \equiv F_2(y) - F_1(y)$, their Method 5 tests $H_{0y}\colon \theta_y \leq 0$ for each $y \in \mathbb{R}$, using order statistics to achieve strong control of finite-sample FWER (see their Theorem 9). Such multiple testing has clear economic significance. For example, let $F_2(\cdot)$ and $F_1(\cdot)$ be the wage CDFs for individuals with and without a high-school degree, respectively, in dollars per hour. Then H_{0y} means individuals without a degree have at least as high probability of wage below $y per hour as individuals with a degree. The multiple testing procedure lets us test such comparisons across all y while controlling the probability of at least one false rejection. Reversing the direction of inequality, rejecting $H_{0y}^{\geq}\colon \theta_y \geq 0$ provides stronger evidence in favor of $F_2(y) < F_1(y)$, meaning that individuals with a degree are less likely than individuals without a degree to have wage below $y per hour.

Extending to any finite $K \geq 2$ with FWER α, we apply their Method 5 to each of the $K-1$ pairs of consecutive $(x, x+1)$ values with Bonferroni-adjusted level $\alpha/(K-1)$. In the special case $K = 2$, there is no adjustment because $K - 1 = 1$. As discussed in our introduction, the Bonferroni adjustment errs on the conservative side, but not too egregiously because the dependence across X is negative rather than positive. (It is important that we only use the Bonferroni adjustment across X and not across Y, which has an infinite number of points with strong positive dependence.) Theorem 4 formalizes this.

Theorem 4. *Consider the multiple testing procedure that applies Method 5 of Goldman and Kaplan (2018) with level $\alpha/(K-1)$ a total of $K-1$ times: first to $H_{0x,y}$ in (14) for $(x,y) \in \{1\} \times \mathbb{R}$, second to $H_{0x,y}$ for $(x,y) \in \{2\} \times \mathbb{R}$, etc., up to $(x,y) \in \{K-1\} \times \mathbb{R}$. Under Assumption A1 with strictly increasing conditional CDFs $F_x(\cdot)$, this procedure has strong control of finite-sample FWER at level α.*

As in Section 2.4, we can invert the multiple testing procedure into confidence sets for the true set $\mathcal{T} \equiv \{(x,y) : \theta_{x,y} \leq 0\}$ from (11). Again, the outer confidence set collects the (x,y) corresponding to $H_{0x,y}\colon \theta_{x,y} \leq 0$ that are not rejected, whereas the inner confidence set collects the (x,y) corresponding to rejected $H_{0x,y}^{\geq}\colon \theta_{x,y} \geq 0$. The finite-sample FWER control translates to finite-sample coverage probabilities.

Theorem 5. *Consider the inner confidence set $\widehat{\mathcal{CS}}_{inner}$ that collects all pairs of (x, y) for which the reversed null $H^{\geq}_{0x,y}\colon \theta_{x,y} \geq 0$ is rejected by the multiple testing procedure in Theorem 4, and the outer confidence set $\widehat{\mathcal{CS}}_{outer}$ that collects all pairs of (x, y) for which the original null $H_{0x,y}\colon \theta_{x,y} \leq 0$ is not rejected. Under Assumption A1 with strictly increasing conditional CDFs $F_x(\cdot)$, these confidence sets have finite-sample coverage probability:*

$$\mathrm{P}(\widehat{\mathcal{CS}}_{inner} \subseteq \mathcal{T}) \geq 1 - \alpha, \quad \mathrm{P}(\widehat{\mathcal{CS}}_{outer} \supseteq \mathcal{T}) \geq 1 - \alpha.$$

Also like before, if instead of $\mathcal{T}$ in (11) we are interested in its complement $\mathcal{T}^{\complement} \equiv \{(x, y) : \theta_{x,y} > 0\}$, then the outer CS is the complement of the inner CS for $\mathcal{T}$, and the inner CS is the complement of the outer CS for $\mathcal{T}$. That is,

$$\mathrm{P}(\widehat{\mathcal{CS}}^{\complement}_{inner} \supseteq \mathcal{T}^{\complement}) = \mathrm{P}(\widehat{\mathcal{CS}}_{inner} \subseteq \mathcal{T}) \geq 1 - \alpha,$$
$$\mathrm{P}(\widehat{\mathcal{CS}}^{\complement}_{outer} \subseteq \mathcal{T}^{\complement}) = \mathrm{P}(\widehat{\mathcal{CS}}_{outer} \supseteq \mathcal{T}) \geq 1 - \alpha.$$

4 Empirical illustrations

To demonstrate our methodology, we apply it to ordinal and continuous outcome examples in Sections 4.1 and 4.2, respectively. All code is in R (R Core Team, 2023), with help from the `MASS` package (Venables and Ripley, 2002) for multivariate normal random vector generation, and from the packages `ggplot2` (Wickham, 2016), `ggmosaic` (Jeppson and Hofmann, 2023; Jeppson, Hofmann, and Cook, 2023), and `scales` (Wickham, Pedersen, and Seidel, 2023) for plotting. Files to replicate our results are available online.[1]

4.1 Mental health and education

Many studies have examined the relationship between depression and education (e.g., Bauldry, 2015; Cohen, Nussbaum, Weintraub, Nichols, and Yen, 2020; Lee, 2011). The nature of the correlation varies across different investigations. Individuals with more education may pos-

[1] `https://qianjoewu.github.io/research/osm/Wu_Kaplan_OSM.zip`

sess better mental health management skills, but they may also more readily recognize and report their psychological states. Further, the relationship is potentially different when (for example) comparing high-school dropouts and graduates than when comparing bachelor's and graduate degree holders, and the relationship may be different for severe depression than for milder forms. Our method can detect any such patterns because it is nonparametric and does not impose any restrictions. We assess evidence of where the level of depression is stochastically decreasing in education, i.e., where mental health is improving.

We use the following variables from the publicly available NHIS 2022 data (National Center for Health Statistics, 2022). Depression variable Y (`PHQCAT_A`) summarizes the eight-item Patient Health Questionnaire depression scale (PHQ-8) into four categories: "none/minimal," "mild," "moderate," and "severe," respectively coded as 1, 2, 3, and 4. Education variable X (recoded from `EDUCP_A`) has five categories: "no high school degree," "HS degree only," "some college" (including associate's degree), "bachelor's degree," and "graduate degree" (master's, professional, or doctoral), respectively coded as 1, 2, 3, 4, and 5. We restrict the age range to 30–64 years old for a fair comparison of people with various levels of education including graduate degrees, and to concentrate on those who are eligible for the workforce. For simplicity, observations with missing values are dropped and weights are not used.

Figure 1 visualizes the data on depression and education in a mosaic plot. Each column corresponds to a category of X, and its width is proportional to the number of observations in that category. For example, the HS column is significantly wider than the no-HS column, showing that more individuals in the data have (only) a high-school degree than do not have a high-school degree. Each cell's area is proportional to the sample proportion of individuals with that particular value of (X, Y); the text label shows that proportion as a percentage, along with the corresponding number of observations in that cell. Because the joint probability of $(X, Y) = (x, y)$ is the product of the marginal $X = x$ probability and the conditional Y probability given $X = x$, within a column corresponding to $X = x$,

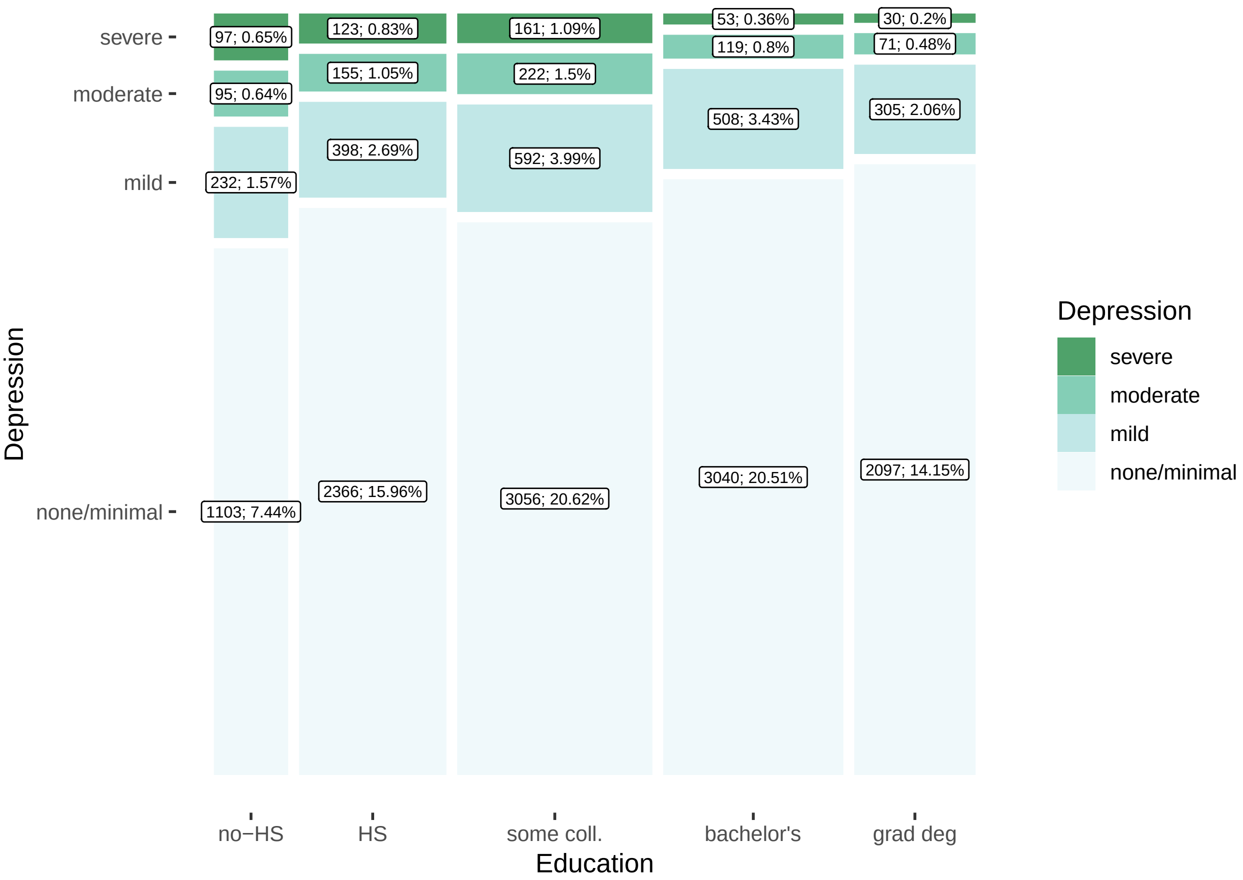

Figure 1: Mosaic plot of depression and education.

each cell's height is proportional to the proportion of observations with the corresponding y value within the $X_i = x$ subsample. These conditional probabilities are scaled such that the full height of each column is probability one (or 100%). For example, in the HS column, the "none/minimal" cell's height is over half the total column height, meaning that among individuals in the data who have (only) a high-school degree, over half have depression level "none/minimal." This further implies that the breaks between cells within a column show the conditional CDF values. For example, in the no-HS column, the top of the "mild" cell is around 7/8 between the bottom and top of the column, indicating the empirical conditional CDF is around $\hat{F}_1(2) \approx 7/8$. For each level y, the mosaic plot shows the data sample follows the pattern $\hat{F}_1(y) \leq \hat{F}_2(y)$ and $\hat{F}_3(y) \leq \hat{F}_4(y) \leq \hat{F}_5(y)$ (i.e., generally lower depression at higher education), but $\hat{F}_2(y) \geq \hat{F}_3(y)$. However, we wish to learn not only about the sample but about the population relationship. Our methods help assess the strength of the evidence that these patterns hold in the population.

Table 1: Test statistics for depression versus education ($\hat{c}_{0.05} = 2.60$).

	Education category x			
Depression level y	1 (vs. 2)	2 (vs. 3)	3 (vs. 4)	4 (vs. 5)
1: none/minimal	**4.04**	-1.94	**6.38**	2.12
2: mild (or below)	**3.45**	-0.52	**8.47**	1.13
3: moderate (or below)	**3.21**	0.10	**7.05**	0.78

True set: points where mental health is better (lower depression) at next-higher education level, $\{(x, y) : F_x(y) \leq F_{x+1}(y)\}$. Gray shading: outer confidence set. Bold: inner confidence set. Confidence level 95%. Critical value computed using $N = 100{,}000$ random draws.

Table 1 shows our inference results, which can be interpreted as follows. The numbers are the t-statistics for testing the null hypotheses $H_{0x,y}\colon F_x(y) \geq F_{x+1}(y)$. For example, the bottom-left number in the table is for $x = 1$ and $y = 3$, corresponding to the t-statistic for the null hypothesis that $F_1(3) \geq F_2(3)$, i.e., that the population probability of having moderate or below depression ($Y \leq 3$) is higher among individuals with no high school degree ($x = 1$) than among individuals who have a high school degree only ($x = 2$). The corresponding t-statistic $\hat{t}_{x,y}(0)$ from (7) is $\hat{t}_{2,1}(0) = (\hat{F}_2(3) - \hat{F}_1(3))/\hat{s}_{2,1} = 3.21$, which is above $\hat{c}_{0.05} = 2.60$ (hence bold **3.21** in the table), so we reject $F_1(3) \geq F_2(3)$ in favor of $F_1(3) < F_2(3)$. Because there are only $J = 4$ categories of Y, we can also interpret the conclusion $F_1(3) < F_2(3)$ as $\mathrm{P}(Y = 4 \mid X = 1) > \mathrm{P}(Y = 4 \mid X = 2)$, i.e., a higher probability of severe depression in the no-high-school subpopulation than in the high-school-only subpopulation. More generally in the table, a t-statistic is positive when $\hat{F}_x(y) < \hat{F}_{x+1}(y)$, indicating a larger proportion of individuals with depression level y or below in the $X_i = x + 1$ subsample than in the $X_i = x$ subsample, generally indicating less depression (better mental health) at the higher education level $x + 1$ than at x. If the t-statistic is positive and above the critical value, then there is strong evidence that this is true in the population, too. Consequently, when $\hat{t}_{x,y}(0) > \hat{c}_{0.05}$, that (x, y) is included in the 95% inner confidence set, which collects the points with strong evidence of depression decreasing with education; such t-statistics are in **bold**.

The 95% outer confidence set includes (x, y) as long as there is no strong evidence in the opposite direction, i.e., as long as $\hat{t}_{x,y}(0) > -\hat{c}_{0.05}$; such t-statistics are shaded gray. Note the $\hat{t}_{x,y}(0) > \hat{c}_{0.05}$ are thus **bold and shaded** because the outer confidence set contains the inner confidence set.

Table 1 shows strong evidence of a general pattern of lower depression with higher education, but with exceptions. Half of the t-statistics exceed our 5% FWER critical value, collectively forming a relatively large 95% inner confidence set for the true set of points $\{(x, y) : F_x(y) \leq F_{x+1}(y)\}$ where mental health is better (lower depression) at the higher education level. This includes all depression levels y for the education level comparisons of no-high-school with high-school-only and of some college with bachelor's degree. The outer confidence set includes all cells, including some with t-statistics that are negative but not quite negative enough for our multiple testing procedure to reject in the opposite direction. Thus, the empirical evidence is consistent with depression stochastically decreasing with education, but our results show that the strength of evidence varies considerably by education level.

Overall, Table 1 provides much richer results than simply saying we do not reject the global null hypothesis of stochastic monotonicity.

4.2 Earnings and education

We assess the evidence for individual earnings stochastically increasing with education using the following variables from the publicly available CPS ASEC 2024 data (The U.S. Census Bureau, 2024). The Y variable (`PEARNVAL`) represents individual annual earnings, which is continuous and ranges from $-\$9999$ to \$2,099,999. For plots, we focus on Y values between 0 and 1,000,000, which covers more than 99.9% of the data. The education variable X (recoded from `A_HGA`) has five categories: "no high school degree," "HS degree only," "some college" (including associate's degree), "bachelor's degree," and "graduate degree" (master's, professional, or doctoral), which we code as values 1, 2, 3, 4, and 5, respectively.

We restrict the age range to 30–64 years old for a fair comparison of people with various levels of education including graduate degrees, and to concentrate on those who do not face significant age-related barriers to employment. For simplicity, observations with missing values are dropped and weights are not used.

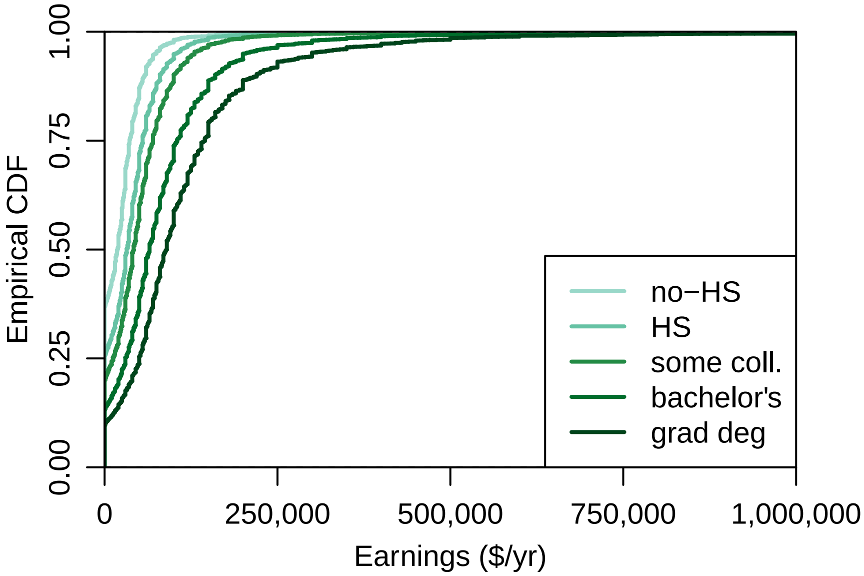

Figure 2: CDF plot of earnings by different education categories.

Figure 2 presents empirical CDFs of individual earnings across the five education categories described above. The graph shows that the empirical CDF curves are ordered such that $\hat{F}_5(\cdot) \leq \hat{F}_4(\cdot) \leq \hat{F}_3(\cdot) \leq \hat{F}_2(\cdot) \leq \hat{F}_1(\cdot)$, indicating higher educational attainment is associated with higher earnings. However, this stochastic monotonicity in the sample does not necessarily imply stochastic monotonicity in the population. Our methods help quantify our uncertainty about the population relationship.

Figure 3 shows our inner and outer confidence sets. We use a 95% confidence level for each, so given Theorem 5, our confidence sets have (at least) 95% coverage not only asymptotically but in finite samples. The true set contains points where the population conditional CDFs are consistent with earnings stochastically increasing in education, $F_{x+1}(y) \leq F_x(y)$. The inner confidence set consists of the dark gray regions, which provide a conservative "estimate" of this true set, in the sense that there is a 95% ex ante (frequentist) probability

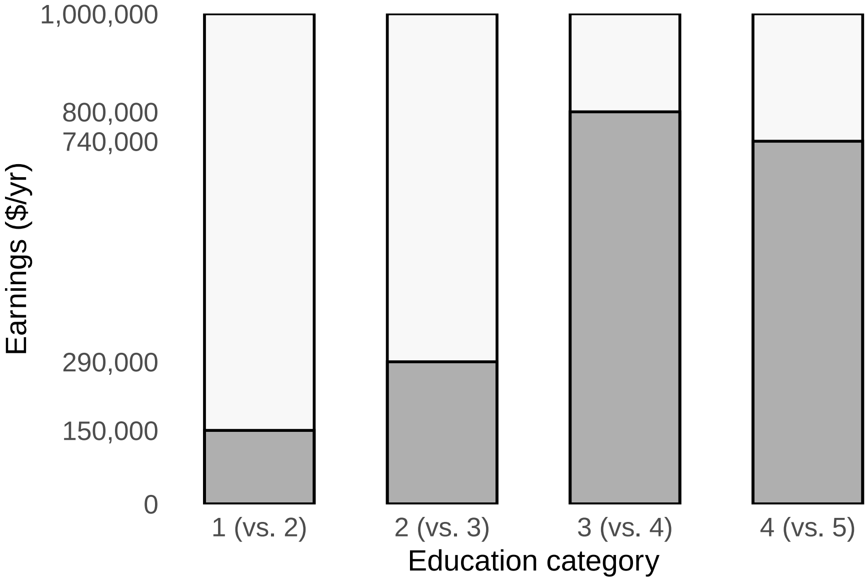

Figure 3: Inner (dark gray) and outer (light gray) 95% confidence sets.

of sampling a dataset for which the inner confidence set is contained within the true set. That is, the dark gray regions show where there is especially strong evidence supporting stochastic increasingness. The outer confidence set includes the dark gray regions as well as the light gray regions, which here are as big as possible. That is, there is no point with strong evidence against stochastic increasingness. Considering the y-axis labels, the inner confidence set is also large. It includes all y values between \$0/yr and \$150,000/yr for all education levels, and even much higher y values at higher education levels. In contrast to the mental health results in Section 4.1, here there is much strong evidence for stochastically increasing earnings at each education level.

In sum, Figure 3 provides more granular and precise evidence supporting the stochastically monotonic relationship between earnings and education, compared to a global test that would merely "not reject" the null hypothesis of stochastic increasingness.

5 Simulations

This section presents simulations of finite-sample FWER for our asymptotically justified multiple testing procedure when Y is ordinal or discrete. We use a variety of sample sizes and numbers of Y and X categories. We use the "least favorable" configuration that maximizes FWER by setting all null hypothesis inequalities to be binding: $\theta_{x,y} = 0$ for all (x, y). We provide code in R (R Core Team, 2023) to replicate our results.[2]

The simulation proceeds as follows. For the data generating process, Y and X respectively have J and K categories. The conditional distribution of Y is uniform across the J categories, so the conditional CDF is $F_x(y) = y/J$ for each (x, y) and all $\theta_{x,y} \equiv F_{x+1}(y) - F_x(y) = y/J - y/J = 0$. After randomly drawing a dataset having n_x observations with $X_i = x$ for each x (so total sample size $n = Kn_x$), Method 1 is run, with the critical value based on N simulations, and the results are stored. This is repeated 1000 times. Because all $H_{0x,y}$ are true, the simulated FWER is the proportion of simulated datasets for which at least one $H_{0x,y}$ is rejected. Besides FWER, we report the full range (across all simulated datasets) of simulated critical values c_α for each case.

Table 2 presents the simulation results. They are divided into three sections corresponding to the different (J, K), but results are qualitatively similar in each section, showing the following patterns. First, FWER is somewhat above $\alpha = 0.05$ with the smallest sample size, closer to 0.10, suggesting that (as usual) results from smaller samples should be interpreted more cautiously. Second, as n_x increases, FWER approaches the nominal α. This reflects our procedure's exact asymptotic FWER in this least favorable configuration; FWER would be lower in other cases. Third, compared to $N = 1000$, computing the critical value using $N = 10{,}000$ draws improves stability somewhat (tighter range of critical values) but improves FWER accuracy only very slightly, so in practice we suggest using $N = 1000$ as the default.

[2] `https://qianjoewu.github.io/research/osm/Wu_Kaplan_OSM.zip`

Table 2: Simulation results.

J	K	n_x	N	α	c_α	FWER
4	4	20	1000	0.05	$[2.32, 2.59]$	0.098
4	4	100	1000	0.05	$[2.34, 2.58]$	0.068
4	4	1000	1000	0.05	$[2.36, 2.57]$	0.060
4	4	10,000	1000	0.05	$[2.36, 2.55]$	0.045
4	4	10,000	10,000	0.05	$[2.43, 2.52]$	0.047
4	4	10,000	10,000	0.10	$[2.18, 2.25]$	0.112
4	4	10,000	10,000	0.01	$[2.94, 3.08]$	0.010
6	5	20	1000	0.05	$[2.59, 2.84]$	0.092
6	5	100	1000	0.05	$[2.55, 2.81]$	0.075
6	5	1000	1000	0.05	$[2.57, 2.79]$	0.068
6	5	10,000	1000	0.05	$[2.60, 2.79]$	0.053
6	5	10,000	10,000	0.05	$[2.70, 2.77]$	0.047
6	5	10,000	10,000	0.10	$[2.45, 2.50]$	0.105
6	5	10,000	10,000	0.01	$[3.18, 3.32]$	0.007
8	10	20	1000	0.05	$[2.92, 3.18]$	0.099
8	10	100	1000	0.05	$[2.95, 3.18]$	0.074
8	10	1000	1000	0.05	$[2.94, 3.20]$	0.056
8	10	10,000	1000	0.05	$[2.95, 3.18]$	0.046
8	10	10,000	10,000	0.05	$[3.04, 3.12]$	0.046
8	10	10,000	10,000	0.10	$[2.82, 2.88]$	0.090
8	10	10,000	10,000	0.01	$[3.47, 3.61]$	0.007

6 Conclusion

We have proposed multiple testing procedures and confidence sets to provide a richer assessment of stochastic monotonicity, by separately considering each conditional CDF inequality that together comprise stochastic monotonicity, for continuous, discrete, or ordinal outcomes. In future work, to complement our frequentist confidence sets, we plan to develop Bayesian credible sets and assess their frequentist properties. For the related *joint* testing problem, the frequentist test is much more conservative than a Bayesian test even asymptotically (e.g., Kaplan and Zhuo, 2021; Kline, 2011), but this does not translate directly to multiple testing and related confidence sets. Deriving multiple testing procedures for continuous X would also be valuable.

Supplementary Materials

We provide R code to replicate all simulation and empirical results.

Acknowledgments

Thanks to the following for their helpful feedback: Alyssa Carlson, Zack Miller, Shawn Ni, journal reviewers, and participants in the Asian Meeting of the Econometric Society (2022, CUHK/virtual), Missouri Valley Economic Association annual conference (2021, virtual), Economics Graduate Student Conference (2022, Washington University in St. Louis), University of Missouri Economics Research Workshop, Chinese Economists Society annual conference (2022, Guizhou/virtual), and Midwest Econometrics Group (2022, Michigan State).

Disclosure Statement

We have no competing interests to declare.

References

Armstrong, Timothy B. and Shu Shen. 2015. "Inference on Optimal Treatment Assignments." Working paper available at `https://tbarmstr.github.io/`.

Bauldry, Shawn. 2015. "Variation in the Protective Effect of Higher Education against Depression." *Society and Mental Health* 5 (2):145–161. URL `https://doi.org/10.1177/2156869314564399`.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society: Series B* 57 (1):289–300. URL `https://www.jstor.org/stable/2346101`.

Canay, Ivan A. and Azeem M. Shaikh. 2017. "Practical and Theoretical Advances in Inference for Partially Identified Models." In *Advances in Economics and Econometrics: Eleventh World Congress*, *Econometric Society Monographs*, vol. 2, edited by Ariel Pakes, Bo Honoré, Larry Samuelson, and Monika Piazzesi, chap. 9. Cambridge: Cambridge University Press, 271–306. URL `https://doi.org/10.1017/9781108227223.009`.

Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov. 2021. "On Cross-Validated Lasso in High Dimensions." *Annals of Statistics* 49 (3):1300–1317. URL `https://doi.org/10.1214/20-AOS2000`.

Cohen, Alison K., Juliet Nussbaum, Miranda L. Ritterman Weintraub, Chloe R. Nichols, and Irene H. Yen. 2020. "Association of Adult Depression With Educational Attainment, Aspirations, and Expectations." *Preventing Chronic Disease* 17 (94). URL https://doi.org/10.5888/pcd17.200098.

Davidson, Russell and Jean-Yves Duclos. 2013. "Testing for Restricted Stochastic Dominance." *Econometric Reviews* 32 (1):84–125. URL https://doi.org/10.1080/07474938.2012.690332.

Goldman, Matt and David M. Kaplan. 2018. "Comparing distributions by multiple testing across quantiles or CDF values." *Journal of Econometrics* 206 (1):143–166. URL https://doi.org/10.1016/j.jeconom.2018.04.003.

Jeppson, Haley and Heike Hofmann. 2023. "Generalized Mosaic Plots in the ggplot2 Framework." *The R Journal* 14 (4):50–78. URL https://doi.org/10.32614/RJ-2023-013.

Jeppson, Haley, Heike Hofmann, and Dianne H. Cook. 2023. *ggmosaic: Mosaic Plots in the 'ggplot2' Framework*. URL https://haleyjeppson.github.io/ggmosaic/,https://github.com/haleyjeppson/ggmosaic. R package version 0.3.4.

Kaplan, David M. 2024. "Inference on Consensus Ranking of Distributions." *Journal of Business & Economic Statistics* 42 (3):839–850. URL https://doi.org/10.1080/07350015.2023.2252040.

Kaplan, David M. and Wei Zhao. 2023. "Comparing latent inequality with ordinal data." *Econometrics Journal* 26 (2):189–214. URL https://doi.org/10.1093/ectj/utac030.

Kaplan, David M. and Longhao Zhuo. 2021. "Frequentist properties of Bayesian inequality tests." *Journal of Econometrics* 221 (1):312–336. URL https://doi.org/10.1016/j.jeconom.2020.05.015.

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 (9):3131–3141. URL https://doi.org/10.1016/j.jspi.2011.03.034.

Lee, Jinkook. 2011. "Pathways from Education to Depression." *Journal of Cross-Cultural Gerontology* 26 (2):121–135. URL https://doi.org/10.1007/s10823-011-9142-1.

Lee, Sokbae, Oliver Linton, and Yoon-Jae Whang. 2009. "Testing for Stochastic Monotonicity." *Econometrica* 77 (2):585–602. URL https://www.jstor.org/stable/40263876.

Lehmann, E. L. and Joseph P. Romano. 2005a. "Generalizations of the Familywise Error Rate." *Annals of Statistics* 33 (3):1138–1154. URL https://projecteuclid.org/euclid.aos/1120224098.

———. 2005b. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 3rd ed. URL https://books.google.com/books?id=Y7vSVW3ebSwC.

Manski, Charles F. 1997. "Monotone Treatment Response." *Econometrica* 65 (6):1311–1334. URL https://doi.org/10.2307/2171738.

National Center for Health Statistics. 2022. "National Health Interview Survey." https://www.cdc.gov/nchs/nhis/documentation/2022-nhis.html. Public-use data file and documentation, accessed 2024.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. 2010. "Multiple testing." In *The New Palgrave Dictionary of Economics*, edited by Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan, online ed., 1–5. URL https://doi.org/10.1057/

9780230226203.3826.

Seo, Juwon. 2018. “Tests of stochastic monotonicity with improved power.” *Journal of Econometrics* 207 (1):53–70. URL https://doi.org/10.1016/j.jeconom.2018.04.004.

Small, Dylan S., Zhiqiang Tan, Roland R. Ramsahai, Scott A. Lorch, and M. Alan Brookhart. 2017. “Instrumental Variable Estimation with a Stochastic Monotonicity Assumption.” *Statistical Science* 32 (4):561–579. URL https://doi.org/10.1214/17-STS623.

The U.S. Census Bureau. 2024. “Annual Social and Economic Supplements.” https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.2024.html#list-tab-165711867. Public-use data file and documentation, accessed 2024.

van der Vaart, Aad W. 1998. *Asymptotic Statistics*. Cambridge: Cambridge University Press. URL https://books.google.com/books?id=UEuQEM5RjWgC.

Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S*. New York: Springer, fourth ed. URL https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL https://ggplot2.tidyverse.org.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *scales: Scale Functions for Visualization*. URL https://scales.r-lib.org. R package version 1.3.0, https://github.com/r-lib/scales.

Zhao, Wei and David M. Kaplan. 2024. “Multiple Testing of a Function’s Monotonicity.” Working paper available at https://kaplandm.github.io/.

A Proofs

A.1 Proof of Lemma 1

Proof. Under Assumption A1 and the central limit theorem (e.g., van der Vaart, 1998, Prop. 2.27), as $n_x \to \infty$,

$$\sqrt{n_x}(\hat{F}_x(y) - F_x(y)) \xrightarrow{d} \mathrm{N}\big(0, F_x(y)[1 - F_x(y)]\big). \tag{15}$$

Let $\hat{\boldsymbol{A}}$ be the estimator of vector $\boldsymbol{A}$ that contains all conditional CDFs,

$$\boldsymbol{A} \equiv (F_1(1), F_1(2), \ldots, F_1(J-1), \ldots, F_K(1), F_K(2), \ldots, F_K(J-1))'. \tag{16}$$

Under Assumption A1, by the continuous mapping theorem (CMT) (e.g., van der Vaart, 1998, Thm. 2.3),

$$\sqrt{n_1}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \underline{\boldsymbol{V}}), \tag{17}$$

where $\boldsymbol{0}$ is a vector of zeros and $\underline{\boldsymbol{V}}$ is the $(JK-K) \times (JK-K)$ block diagonal matrix

$$\underline{\boldsymbol{V}} \equiv \begin{bmatrix} \underline{\boldsymbol{V}}^1 & 0 & \cdots & 0 \\ 0 & \underline{\boldsymbol{V}}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \underline{\boldsymbol{V}}^K \end{bmatrix} \tag{18}$$

where each $\underline{\boldsymbol{V}}^k$ is a $(J-1) \times (J-1)$ matrix with entry in row i, column j denoted

$$V^k_{ij} = \gamma_x F_x(\min\{i,j\})[1 - F_x(\max\{i,j\})].$$

We apply the delta method (e.g., van der Vaart, 1998, Thm. 3.1) to derive the asymptotic distribution of the full vector $\hat{\boldsymbol{\theta}}$ for testing the stochastic monotonicity inequalities. For $x \in \{1, 2, \ldots, K-1\}$ and $y \in \{1, 2, \ldots, J-1\}$,

$$\hat{\boldsymbol{\theta}} \equiv (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_{J-1})', \tag{19}$$

where each $\hat{\boldsymbol{\theta}}_j \equiv (\hat{\theta}_{K-1,j}, \hat{\theta}_{K-2,j}, \ldots, \hat{\theta}_{1,j})$ and $\hat{\theta}_{x,y} \equiv \hat{F}_{x+1}(y) - \hat{F}_x(y)$ is the estimator of $\theta_{x,y}$ defined in (4). Then,

$$\sqrt{n_1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{W}}), \quad \underline{\boldsymbol{W}} \equiv \underline{\boldsymbol{H}}' \underline{\boldsymbol{V}} \underline{\boldsymbol{H}}, \tag{20}$$

where $\underline{\boldsymbol{H}} = \frac{\partial}{\partial \boldsymbol{A}} \boldsymbol{\theta}'$ is the partial derivative of the vector $\boldsymbol{\theta}'$ with respect to $\boldsymbol{A}$ in (16), and $\underline{\boldsymbol{V}}$ is from (18). Note each element of $\underline{\boldsymbol{H}}$ is either -1, 0, or 1; more specifically, within each column of $\underline{\boldsymbol{H}}$ (the derivative of a particular $\theta_{x,y}$ with respect to $\boldsymbol{A}$), one element is -1, one element is 1, and the rest equal zero.

From (20), for each $\hat{\theta}_{x,y}$, using $n_1/n_x \to \gamma_x$ from A1,

$$\begin{aligned}
\sqrt{n_1}(\hat{\theta}_{x,y} - \theta_{x,y}) &= \sqrt{n_1}\{[\hat{F}_{x+1}(y) - \hat{F}_x(y)] - [F_{x+1}(y) - F_x(y)]\} \\
&= \overbrace{\sqrt{n_1/n_{x+1}}}^{\to\sqrt{\gamma_{x+1}}} \overbrace{\sqrt{n_{x+1}}[\hat{F}_{x+1}(y) - F_{x+1}(y)]}^{\text{use (15)}} - \overbrace{\sqrt{n_1/n_x}}^{\to\sqrt{\gamma_x}} \overbrace{\sqrt{n_x}[\hat{F}_x(y) - F_x(y)]}^{\text{use (15)}} \\
&\xrightarrow{d} \mathrm{N}(0, \sigma_{x,y}^2), \\
\sigma_{x,y} &\equiv \sqrt{\gamma_x F_x(y)[1 - F_x(y)] + \gamma_{x+1} F_{x+1}(y)[1 - F_{x+1}(y)]},
\end{aligned} \tag{21}$$

where the variances are summed because the subsamples for $X_i = x$ and $X_i = x + 1$ are independent given A1, and if $W \perp\!\!\!\perp Z$ then $\mathrm{Var}(W + Z) = \mathrm{Var}(W) + \mathrm{Var}(Z)$.

For the standard error of $\hat{\theta}_{x,y}$, which is asymptotically $\sigma_{x,y}/\sqrt{n_1}$ given (21), we use the estimator

$$\hat{s}_{x,y} \equiv \sqrt{\frac{(n_1/n_x)\hat{F}_x(y)(1 - \hat{F}_x(y)) + (n_1/n_{x+1})\hat{F}_{x+1}(y)(1 - \hat{F}_{x+1}(y))}{n_1}}. \tag{22}$$

By the consistency of the conditional CDF estimators and the CMT, as $n_1 \to \infty$,

$$\begin{aligned}
\sqrt{n_1}\hat{s}_{x,y} &= \sqrt{(n_1/n_x)\hat{F}_x(y)(1 - \hat{F}_x(y)) + (n_1/n_{x+1})\hat{F}_{x+1}(y)(1 - \hat{F}_{x+1}(y))} \\
&\xrightarrow{p} \sqrt{\gamma_x F_x(y)(1 - F_x(y)) + \gamma_{x+1} F_{x+1}(y)(1 - F_{x+1}(y))} = \sigma_{x,y},
\end{aligned} \tag{23}$$

the asymptotic standard deviation in (21). Informally, the standard error estimator $\hat{s}_{x,y}$ is "consistent," meaning formally that $\sqrt{n_1}\hat{s}_{x,y} \xrightarrow{p} \sigma_{x,y}$.

Substituting (22) into (7),

$$\hat{t}_{x,y}(\theta_{x,y}) = \frac{\hat{\theta}_{x,y} - \theta_{x,y}}{\hat{s}_{x,y}} = \frac{\sqrt{n_1}(\hat{\theta}_{x,y} - \theta_{x,y})}{\sqrt{(n_1/n_x)\hat{F}_x(y)(1 - \hat{F}_x(y)) + (n_1/n_{x+1})\hat{F}_{x+1}(y)(1 - \hat{F}_{x+1}(y))}}.$$

Thus, we can write the vector of properly centered t-statistics as $\hat{\boldsymbol{t}} = \sqrt{n_1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\hat{\underline{\boldsymbol{S}}}$, where $\sqrt{n_1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is the left-hand side of (20), and $\hat{\underline{\boldsymbol{S}}}$ is a diagonal matrix having elements $1/(\hat{s}_{x,y}\sqrt{n_1})$ matching the corresponding (x, y) from the $\boldsymbol{\theta}$ vector; equivalently, the row i, column j element of the matrix $\hat{\underline{\boldsymbol{S}}}$ is $\mathbb{1}\{i = j\}\hat{W}_{ij}$. By (23), $\hat{\underline{\boldsymbol{S}}} \xrightarrow{p} \underline{\boldsymbol{S}}$, the diagonal matrix with corresponding (x, y) elements $1/\sigma_{x,y}$,

$$\underline{\boldsymbol{S}} \equiv \begin{bmatrix} 1/\sigma_{K-1,1} & 0 & \cdots & 0 \\ 0 & 1/\sigma_{K-2,1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_{1,J-1} \end{bmatrix}. \tag{24}$$

By (20) and CMT, $\hat{\boldsymbol{t}} \equiv \sqrt{n_1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\hat{\underline{\boldsymbol{S}}} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{S}}'\underline{\boldsymbol{W}}\,\underline{\boldsymbol{S}})$. □

A.2 Proof of Theorem 2

Proof. When testing original hypotheses $H_{0x,y}\colon F_{x+1}(y) - F_x(y) \le 0$, the set of true hypotheses is $\mathcal{T} \equiv \{(x, y) : F_{x+1}(y) - F_x(y) \le 0\}$. For any true $H_{0x,y}$,

$$\hat{t}_{x,y}(0) = \frac{\hat{F}_{x+1}(y) - \hat{F}_x(y)}{\hat{s}_{x,y}} \le \frac{\hat{F}_{x+1}(y) - \hat{F}_x(y) - \overbrace{[F_{x+1}(y) - F_x(y)]}^{\le 0 \text{ given true } H_{0x,y}}}{\hat{s}_{x,y}} \equiv \hat{t}_{x,y}(\overbrace{F_{x+1}(y) - F_x(y)}^{\theta_{x,y}}). \tag{25}$$

Thus,

$$\begin{aligned} \mathrm{FWER} &\equiv \mathrm{P}(\text{reject } H_{0x,y} \text{ for any } (x, y) \in \mathcal{T}) \\ &= \mathrm{P}(\hat{t}_{x,y}(0) > \hat{c}_\alpha \text{ for any } (x, y) \in \mathcal{T}) \\ &= \mathrm{P}(\max_{(x,y)\in\mathcal{T}} \overbrace{\hat{t}_{x,y}(0)}^{\text{use (25)}} > \hat{c}_\alpha) \end{aligned}$$

$$\begin{aligned}
&\leq \mathrm{P}(\max_{(x,y)\in\mathcal{T}} \hat{t}_{x,y}(\theta_{x,y}) > \hat{c}_{\alpha}) \\
&\leq \mathrm{P}(\max_{\text{any }(x,y)} \hat{t}_{x,y}(\theta_{x,y}) > \hat{c}_{\alpha}) \\
&\to \mathrm{P}(\max_{\text{any }(x,y)} \overbrace{t_{x,y}}^{\text{from Lemma 1}} > c_{\alpha}) \\
&= \alpha,
\end{aligned} \tag{26}$$

where the second inequality holds because $\mathcal{T} \subseteq \{1,\ldots,K-1\}\times\{1,\ldots,J-1\}$ ("any" (x,y)), and the convergence holds by Lemma 1 and the CMT (because max is continuous), along with $\hat{c}_{\alpha} \xrightarrow{p} c_{\alpha}$, where $\hat{c}_{\alpha}$ was defined in (10) as the $(1-\alpha)$-quantile of the max of a random vector following the $\mathrm{N}(\mathbf{0}, \hat{\underline{\boldsymbol{\Sigma}}})$ distribution (with estimator $\hat{\underline{\boldsymbol{\Sigma}}}$) and c_{α} was defined in (9) as the $(1-\alpha)$-quantile of the max of $\mathrm{N}(\mathbf{0}, \underline{\boldsymbol{\Sigma}})$ (with true $\underline{\boldsymbol{\Sigma}}$). This last part follows because $\hat{\underline{\boldsymbol{\Sigma}}} \xrightarrow{p} \underline{\boldsymbol{\Sigma}}$ (given A1 and CMT) and applying the CMT (given that the max of a normal vector has a continuous, strictly increasing distribution function and thus continuous quantile function).

The asymptotic FWER bound for testing the reversed hypotheses $H^{\geq}_{0x,y}\colon F_{x+1}(y) - F_x(y) \geq 0$ follows essentially the same derivation with the same reason for each step below. Now defining $\mathcal{T}^{\geq} \equiv \{(x,y) : F_{x+1}(y) - F_x(y) \geq 0\}$,

$$\begin{aligned}
\text{FWER} &\equiv \mathrm{P}(\text{reject } H^{\geq}_{0x,y} \text{ for any } (x,y) \in \mathcal{T}^{\geq}) \\
&= \mathrm{P}(\hat{t}_{x,y}(0) < -\hat{c}_{\alpha} \text{ for any } (x,y) \in \mathcal{T}^{\geq}) \\
&= \mathrm{P}(\min_{(x,y)\in\mathcal{T}^{\geq}} \hat{t}_{x,y}(0) < -\hat{c}_{\alpha}) \\
&\leq \mathrm{P}(\min_{(x,y)\in\mathcal{T}^{\geq}} \hat{t}_{x,y}(\theta_{x,y}) < -\hat{c}_{\alpha}) \\
&\leq \mathrm{P}(\min_{\text{any }(x,y)} \hat{t}_{x,y}(\theta_{x,y}) < -\hat{c}_{\alpha}) \\
&\to \mathrm{P}(\min_{\text{any }(x,y)} t_{x,y} < -c_{\alpha}) = \mathrm{P}(-\min_{\text{any }(x,y)} t_{x,y} > c_{\alpha}) = \mathrm{P}(\max_{\text{any }(x,y)} t_{x,y} > c_{\alpha}) = \alpha,
\end{aligned}$$

also using the symmetry of the $\boldsymbol{t} \sim \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{\Sigma}})$ distribution that implies $\max \boldsymbol{t} \overset{d}{=} -\min \boldsymbol{t}$. $\square$

A.3 Proof of Theorem 3

Proof. For the inner CS, we want to derive the probability of the event that the inner CS is contained within $\mathcal{T}$, which is equivalently characterized as "$(x, y) \in \widehat{\mathcal{CS}}_{inner}$ only if $(x, y) \in \mathcal{T}$." With reversed hypotheses $H^{\geq}_{0x,y}\colon \theta_{x,y} \geq 0$ but $\mathcal{T} \equiv \{(x, y) : \theta_{x,y} \leq 0\}$,

$$\begin{aligned}
\mathrm{P}(\widehat{\mathcal{CS}}_{inner} \subseteq \mathcal{T}) &= \mathrm{P}(\text{reject any } H^{\geq}_{0x,y} \text{ only if } (x, y) \in \mathcal{T}) \\
&= \mathrm{P}(\text{reject any } H^{\geq}_{0x,y} \text{ only if } \theta_{x,y} \leq 0) \\
&= 1 - \mathrm{P}(\text{reject any } H_{0x,y} \text{ when } \theta_{x,y} > 0) \\
&\geq 1 - \mathrm{P}(\text{reject any } H_{0x,y} \text{ when } \theta_{x,y} \geq 0) \\
&= 1 - \mathrm{FWER} \geq 1 - \alpha + o(1)
\end{aligned}$$

because FWER $\leq \alpha + o(1)$ by Theorem 2.

The outer CS contains all elements in the true set $\mathcal{T}$ if and only if we do not reject any original null hypothesis $H_{0x,y}\colon \theta_{x,y} \leq 0$ with $(x, y) \in \mathcal{T}$. Similar to the derivation above, the coverage probability is

$$\begin{aligned}
\mathrm{P}(\mathcal{T} \subseteq \widehat{\mathcal{CS}}_{outer}) &= \mathrm{P}(\text{no } H_{0x,y} \text{ rejected with } (x, y) \in \mathcal{T}\ ) \\
&= 1 - \mathrm{P}(\text{reject any } H_{0x,y} \text{ with } \theta_{x,y} \leq 0) \\
&= 1 - \mathrm{FWER} \\
&\geq 1 - \alpha + o(1)
\end{aligned}$$

again because FWER $\leq \alpha + o(1)$ by Theorem 2. □

A.4 Proof of Theorem 4

Proof. We apply a typical Bonferroni strategy given that Method 5 of Goldman and Kaplan (2018) controls finite-sample FWER at adjusted level $\alpha/(K - 1)$ for any given x. The finite-sample FWER is

$$\begin{aligned}
\text{FWER} &\equiv \mathrm{P}(\text{reject any true } H_{0x,y} \text{ over } (x,y) \in \mathcal{X} \times \mathbb{R}) \\
&= \mathrm{P}\left(\bigcup_{x=1}^{K-1} \text{reject any true } H_{0x,y} \text{ over } y \in \mathbb{R} \right) \\
&\le \sum_{x=1}^{K-1} \overbrace{\mathrm{P}(\text{reject any true } H_{0x,y} \text{ over } y \in \mathbb{R})}^{\le \alpha/(K-1) \text{ by Goldman and Kaplan (2018)}} \\
&\le (K-1)\alpha/(K-1) = \alpha. \qquad \square
\end{aligned}$$

A.5 Proof of Theorem 5

Proof. The proof is identical to that of Theorem 3 except the very final inequality because here we have FWER $\le \alpha$ instead of FWER $\le \alpha + o(1)$. Thus, instead of $1 - \text{FWER} \ge 1 - \alpha + o(1)$, the final line becomes $1 - \text{FWER} \ge 1 - \alpha$. □