Finite-Sample Non-Parametric Bounds with an Application to the Causal Effect of Workforce Gender Diversity on Firm Performance

Grace Lordan[1] and Kaveh S. Nobari[1,2]

[1]The Inclusion Initiative, London School of Economics and Political Science, UK
[2]Data Science Institute, London School of Economics and Political Science, UK

September 3, 2025

Abstract

Classical Manski bounds identify average treatment effects under minimal assumptions but, in finite samples, assume that latent conditional expectations are bounded by the sample's own extrema or that the population extrema are known a priori—often untrue in firm-level data with heavy tails. We develop a finite-sample, concentration-driven band (concATE) that replaces that assumption with a Dvoretzky–Kiefer–Wolfowitz tail bound, combines it with delta-method variance, and allocates size via Bonferroni. The band extends to a group-sequential design that controls the family-wise error when the first "significant" diversity threshold is data-chosen. Applied to 945 listed firms (2015 Q2–2022 Q1) concATE shows that senior-level gender diversity raises Tobin's Q once representation exceeds $\approx 30\%$ in growth sectors and $\approx 65\%$ in cyclical sectors.

Keywords: concATE; partial identification; average treatment effect; confidence band; workforce diversity; Tobin's Q; threshold effects
JEL Classification: C21; C14; M14; L25; J16

1 Introduction

Estimating causal effects in settings with partially unobserved counterfactuals is a fundamental challenge in econometrics. Whenever only one of two potential outcomes is observed for each unit, the average treatment effect (ATE) cannot be point-identified without additional assumptions. Recognizing this, a stream of research following Manski's seminal work has developed nonparametric bounds for causal effects under minimal assumptions (Manski, 1990, 2003). The classical Manski bounds make virtually no assumptions beyond knowing the treatment

status and an outcome bound, instead asking: how large or small could the true ATE be, given the data we actually observe? While this worst-case approach guarantees partial identification under arbitrary heterogeneity and certain forms of selection on unobservables, it comes at the cost of wide intervals. In policy settings where the cost of acting on a wrong sign is high, such honesty can be preferable to a potentially misleading precise estimate. However, Manski's bounds have a critical limitation in finite samples: they implicitly assume the unseen counterfactual outcomes lie within known extremes (e.g. the sample minima and maxima or exogenously given bounds). In practice—especially with heavy-tailed outcomes like firm performance—this assumption is often violated. If the true outcome distribution extends beyond the observed range, the traditional bounds can severely undercover the true effect or even become uninformative (infinite) when no credible global bound is available. In short, classical bounds that are valid asymptotically may fail to cover the true ATE in finite samples when outcomes are unbounded. This exposes a methodological gap: how can we conduct inference on partially identified effects without assuming away heavy-tail risks or sacrificing finite-sample validity?

We address this gap by proposing a finite-sample, concentration-driven confidence band for the ATE—henceforth concATE. The concATE methodology replaces Manski's reliance on known outcome bounds with a probabilistic concentration bound that accounts for sampling uncertainty in extreme values. Specifically, we exploit the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Dvoretzky et al., 1956), which provides a tight, distribution-free bound on the maximum discrepancy between the empirical distribution and the true cumulative distribution. By using the DKW inequality, we can guarantee with high finite-sample probability that the empirical range (or other tail statistics) bounds the latent outcome distribution. In essence, instead of assuming the sample extremes equal the population extremes, we allow a margin such that each unobserved tail probability is covered with finite-sample confidence. We then incorporate this "DKW padding" into the estimation of Manski's upper and lower bound components. To account for sampling variability in the observable parts (like the treated and untreated outcome means), we employ standard delta-method approximations. Finally, we combine these elements using Bonferroni's inequality to construct a simultaneous confidence band for the ATE bounds. This concATE band controls the familywise error rate for the entire interval estimate, ensuring that with (for example) 95% confidence the true ATE lies within the band.

Notably, the concATE procedure remains valid under quite general conditions: we require no parametric outcome distribution, only mild tail assumptions and allow for either independent or weakly dependent observations (such as time-series panels) with appropriate mixing conditions. The resulting inference is robust in finite samples, avoiding the need for large-sample approximations or unknown nuisance constants that plague fully nonparametric approaches. By construction, concATE eliminates the strong functional form and ignorability assumptions that conventional regression-based analyses demand, delivering credible inference even when treatment assignment may be endogenous or outcomes are highly non-normal. In contrast to an OLS or panel regression that produces a single point estimate under strict assumptions (Angrist and Pischke, 2009), our approach yields a range of plausible effects consistent with the data and lets the data speak when

identification is weak. In summary, concATE provides a new tool for causal inference under partial observability, offering the transparency of Manski's bounds with greater practical applicability in finite samples.

In addition to its base formulation, our methodology accommodates situations where the parameter of interest is defined only after looking at the data. In particular, we extend concATE to a sequential testing framework to pinpoint ex post a threshold at which the treatment effect becomes nonzero. This extension is motivated by empirical contexts where one expects a non-linear "tipping point" effect rather than a uniform treatment effect. This innovation is especially useful in applications exploring threshold effects, allowing researchers to identify critical values of continuous treatments while rigorously controlling inference error rates.

To demonstrate the utility of our approach, we apply it to the question: Does greater senior-level gender workforce diversity causally improve firm performance? This question has taken on renewed importance as many firms have invested heavily in Diversity, Equity, and Inclusion (DEI) initiatives, yet establishing causality is difficult due to non-random adoption of diversity practices. A rich literature in management and economics has examined links between top management team composition and organizational outcomes. The foundational "upper echelons" theory of Hambrick and Mason (1984) posits that a firm's strategies and performance reflect the backgrounds of its senior executives. Consistent with this view, numerous studies document associations between executive attributes and firm outcomes such as innovation and financial performance. For example, prior research finds that gender-diverse boards tend to exhibit improved internal governance (e.g., better oversight and attendance) although the average impact on firm profitability or market value is mixed (Adams and Ferreira, 2009). A comprehensive meta-analysis by Post and Byron (2015) reports that female board representation is positively related to accounting returns, especially in societies with greater gender parity, but the correlation with market-based performance metrics is weaker. More recent work has begun to address endogeneity in this relationship: Safiullah et al. (2022), analyzing Spain's Gender Equality Act, use GMM techniques and find that while gender-diverse boards outperform on accounting measures, they can underperform on market valuation measures, suggesting investors may respond differently than internal metrics. Similarly, a study of Russian firms by Safiullah et al. (2022) finds that gender-diverse boards are associated with higher profitability and market value, with the benefits particularly pronounced during economic downturns. Beyond gender, other aspects of diversity have been linked to innovation outcomes: Østergaard et al. (2011) show that employee gender and educational diversity positively predict firm innovation, and in a study of London firms, Nathan and Lee (2013) find cultural diversity in management boosts product innovation and entrepreneurship. Field experiments also echo these benefits—Hoogendoorn et al. (2013) conducted a randomized experiment with startup teams and found that gender-balanced teams outperformed male-dominated teams in terms of sales and profits.

An intriguing hypothesis within this literature is the existence of non-linear effects or "critical mass" thresholds in the diversity–performance relationship. Sociologist Rosabeth Kanter's classic work on tokenism (Kanter, 1977, 1987) theorized that women in extreme minority (a "token" few) face marginalization, whereas once a minority group reaches a substantial share of the team, dynamics shift

and their influence grows disproportionately. Kanter's typology categorizes group gender composition into skewed (up to ∼15% women), tilted (∼20–35% women), and balanced (∼40–50% women) categories, proposing that performance benefits might emerge when moving from skewed to tilted or balanced distributions. Subsequent studies have sought empirical evidence of such tipping points. For instance, Torchia et al. (2011) find that having at least three women directors (roughly a critical mass on many boards) is associated with a jump in innovation outputs, consistent with moving beyond token representation. Ali et al. (2011) report an inverted U-shaped relationship between female representation and firm performance in certain contexts, suggesting that the strongest returns may occur at intermediate diversity levels before tapering off. These studies, while suggestive, largely report correlations or rely on linear/quadratic models that may not capture the true causal threshold. Our study contributes to this literature by using a robust, partially identified approach to formally test for causal tipping points. By refraining from imposing a specific functional form, we let the data reveal whether and where increasing diversity has a statistically reliable positive effect on firm value.

Our empirical analysis uses a panel of 945 publicly listed firms observed quarterly from 2015 Q2 to 2022 Q1. We focus on Tobin's Q (the ratio of market value to the replacement cost of assets) as the outcome of interest, which is a standard proxy for a firm's growth opportunities and innovative performance. Originally introduced by Tobin (1969) and later expounded in Tobin's subsequent work (Tobin, 1978), the Q-ratio captures market expectations of future returns—a value above 1 indicates that the firm's market valuation exceeds book value, signaling strong investment incentives (Brainard and Tobin, 1968; Tobin and Brainard, 1976). For each quarter, we define the "treatment" as whether the firm's top management team or board exceeds a given diversity threshold. In separate analyses, we consider thresholds for the percentage of women in senior leadership (e.g., 30%, 40%, 50%, etc.), reflecting the critical mass levels discussed above. We then estimate the nonparametric bounds on the ATE of diversity at each threshold using our concATE procedure. This approach does not assume that firms with different diversity levels are comparable on unobservables; instead, it provides an interval estimate for the possible causal effect, given the observable data, without invoking full identification. In contrast to most prior studies that report point estimates after making identification assumptions, our results will highlight the range of plausible causal impacts of diversity on Tobin's Q, emphasizing what can be learned with minimal assumptions. Our findings yield informative insights. In broad terms, the concATE analysis suggests that senior-level gender diversity has a significantly positive causal effect on Tobin's Q—but only after a certain threshold of representation is achieved. In innovation-driven sectors (such as technology and healthcare, where overall growth opportunities are high), we find that once female representation in leadership surpasses roughly one-third, the lower bound of the ATE becomes positive and the confidence band excludes zero. The estimated effect size grows as diversity increases, with particularly strong gains evident as teams approach gender balance (around 50% female). This provides empirical support for the notion of a "tipping point" around moderate to high diversity levels in dynamic industries. One interpretation is that innovation-oriented firms, facing fast-moving and competitive markets, have strong incentives to harness the bene-

fits of workforce diversity. Such firms may actively invest in inclusive cultures and leadership practices that allow diverse perspectives to be heard and integrated, thereby capturing value from diversity once a basic critical mass is present. By contrast, in more traditional or cyclically oriented industries, the data suggest that a much higher critical mass—on the order of two-thirds female representation—is needed before we detect a reliably positive impact on firm value. Below that level, the confidence bands include zero, indicating we cannot rule out no effect in those sectors. This stringent "tipping point" in traditional industries may reflect a lack of inclusion; when women remain a small minority, they may not experience the psychological safety needed to freely voice their insights or challenge prevailing viewpoints. Alternatively, it is possible that the gains to diversity are lower in this context. Importantly, these conclusions are drawn with rigorous uncertainty quantification. The concATE bands allow us to assert, for example, that at 95% confidence a firm in a growth industry with a gender-balanced leadership enjoys an ATE on Tobin's Q that is positive (bounded away from zero), whereas at lower diversity levels the effect cannot be distinguished from zero. Such results illustrate how our methodological innovation can uncover nuanced causal relationships that might be obscured or misstated by conventional point estimation approaches.

The remainder of the paper is organized as follows. Section 2 formalizes the problem and presents the theoretical framework for nonparametric identification, extending Manski's bounds to our context. Section 3 describes the data, variable construction (particularly the diversity measures), and our estimation procedure in practice. Section 4 details the construction of the finite-sample concATE confidence band and its extension to sequential threshold analysis, including the theoretical guarantees. Section 5 reports results from a Monte Carlo simulation that compares the finite-sample performance of concATE to traditional methods. Section 6 then presents the empirical findings from our panel of firms, highlighting the estimated diversity tipping points and their interpretation. Finally, 7 concludes with a discussion of implications for research and policy, and potential extensions of our framework.

2 Framework

In this section, inspired by the noted shortcomings in causal inference methodologies rigorously discussed by Angrist and Pischke (2009), we extend the theoretical framework of Manski (1990, 2003) to derive nonparametric bounds on the "average" diversity treatment effect.

2.1 Nonparametric Bounds

Let us denote the potential outcomes for firm i in sector j in quarter t by $Y_{ijt}^{(0)}$ and $Y_{ijt}^{(1)}$, corresponding to the scenarios of no diversity efforts (no treatment) and with diversity efforts (treatment), respectively. Regardless of whether firm i actually adopts diversity, $Y_{ijt}^{(0)}$ represents the hypothetical (counterfactual) outcome had the firm not exercised any diversity efforts, and $Y_{ijt}^{(1)}$ represents the outcome if the firm did adopt diversity. In essence, the question we seek to investigate is whether these potential outcomes differ—i.e., whether diversity efforts affect Y_{ijt}.

For simplicity of exposition, assume that $Y_{ijt}^{(k)} \in \mathbb{R}$ for $k \in \{0,1\}$ and define the treatment indicator

$$Z_{ijt}(\tau) = \mathbb{1}\left\{\mathcal{D} \geq \tau\right\}, \tag{1}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, $\mathcal{D}(z_{ijt})$ is a diversity signal, and τ is a threshold chosen by the investigator. Let $\mathbf{X}_{ijt} = (X_{ijt}^1, \ldots, X_{ijt}^p)^\top \in \mathbb{R}^p$ denote a $(p \times 1)$ vector of control variables. Our goal is to learn the conditional treatment effect $Y_{ijt}^{(1)} - Y_{ijt}^{(0)} \mid \mathbf{X}_{ijt}$. Following the notation of Angrist and Pischke (2009), the *observed* outcome can be written in terms of *potential* outcomes as

$$Y_{ijt} = \begin{cases} Y_{ijt}^{(1)}, & \text{if } Z_{ijt} = 1, \\ Y_{ijt}^{(0)}, & \text{if } Z_{ijt} = 0, \end{cases} \tag{2}$$

$$= Y_{ijt}^{(0)} + \left(Y_{ijt}^{(1)} - Y_{ijt}^{(0)}\right) Z_{ijt}. \tag{3}$$

Because only one potential outcome is ever observed for a given firm–quarter (i, j, t), a naïve comparison of conditional means by treatment status is

$$\mathbb{E}\left[Y_{ijt} \mid \mathbf{X}_{ijt}, Z_{ijt} = 1\right] - \mathbb{E}\left[Y_{ijt} \mid \mathbf{X}_{ijt}, Z_{ijt} = 0\right]. \tag{4}$$

Substituting (3) into (4) gives

$$\begin{aligned} \delta(\mathbf{X}) =& \mathbb{E}\left[Y_{ijt} \mid \mathbf{X}_{ijt}, Z_{ijt} = 1\right] - \mathbb{E}\left[Y_{ijt} \mid \mathbf{X}_{ijt}, Z_{ijt} = 0\right] \\ =& \underbrace{\mathbb{E}\left[Y_{ijt}^{(1)} - Y_{ijt}^{(0)} \mid \mathbf{X}_{ijt}, Z_{ijt} = 1\right]}_{\rho(\mathbf{X})} \\ &+ \underbrace{\mathbb{E}\left[Y_{ijt}^{(0)} \mid \mathbf{X}_{ijt}, Z_{ijt} = 1\right] - \mathbb{E}\left[Y_{ijt}^{(0)} \mid \mathbf{X}_{ijt}, Z_{ijt} = 0\right]}_{\mathcal{B}(\mathbf{X})}, \end{aligned} \tag{5}$$

where $\rho(\mathbf{X})$ is the (conditional) treatment effect and $\mathcal{B}(\mathbf{X})$ is the selection bias. As a corollary, the unconditional mean-comparison parameter δ of Angrist and Pischke (2009) is obtained by integrating $\delta(\mathbf{X}_{ijt})$ over the distribution of $\mathbf{X}_{ijt}$. Hence

$$\delta = \mathbb{E}_{\mathbf{X}|Z}\left[\delta(\mathbf{X})\right] = \mathbb{E}\left[Y_{ijt} \mid Z_{ijt} = 1\right] - \mathbb{E}\left[Y_{ijt} \mid Z_{ijt} = 0\right]. \tag{6}$$

The latter may be non-zero because firms that adopt diversity efforts might do so precisely when they face innovation shortfalls, either to signal responsiveness to investors or to diversify their workforce in search of new ideas; in such cases $\mathcal{B}(\mathbf{X}) < 0$. Conversely, if a firm scales up diversity after large innovation gains, aiming to sustain that momentum, then $\mathcal{B}(\mathbf{X}) > 0$.

Manski (1990, 2003) formalise the problem differently. For firms characterised by attributes $\mathbf{X}$, define the difference in expected outcomes as

$$\begin{aligned} \Re(\mathbf{X}) &= \mathbb{E}\left[Y_{ijt}^{(1)} \mid \mathbf{X}_{ijt}\right] - \mathbb{E}\left[Y_{ijt}^{(0)} \mid \mathbf{X}_{ijt}\right] \\ &= \mathbb{E}\left[Y_{ijt}^{(1)} - Y_{ijt}^{(0)} \mid \mathbf{X}_{ijt}\right]. \end{aligned} \tag{7}$$

Using the law of total expectation, each conditional mean in (7) can be decomposed; for example,

$$\begin{aligned}\mathbb{E}\left[Y_{ijt}^{(1)} \mid \mathbf{X}_{ijt}\right] = \mathbb{E}\left[Y_{ijt}^{(1)} \mid \mathbf{X}_{ijt}, Z_{ijt}=1\right] \Pr\left(Z_{ijt}=1 \mid \mathbf{X}_{ijt}\right) \\ + \mathbb{E}\left[Y_{ijt}^{(1)} \mid \mathbf{X}_{ijt}, Z_{ijt}=0\right] \Pr\left(Z_{ijt}=0 \mid \mathbf{X}_{ijt}\right),\end{aligned} \tag{8}$$

and an analogous expression holds for $k=0$.

In the conditional-mean comparison of (5), the term $\mathcal{B}(\mathbf{X})$ captures *selection bias*. Equation (8) makes clear that two latent expectations,

$$\mathbb{E}\left[Y_{ijt}^{(1)} \mid \mathbf{X}_{ijt}, Z_{ijt}=0\right] \quad \text{and} \quad \mathbb{E}\left[Y_{ijt}^{(0)} \mid \mathbf{X}_{ijt}, Z_{ijt}=1\right], \tag{9}$$

are never observed in the data. Put differently, we do not observe the innovation outcome a firm *would have* achieved without diversity efforts when it actually implemented them ($Z_{ijt}=1$), nor the outcome *with* diversity efforts when it did not implement them ($Z_{ijt}=0$).

In both scenarios, one can conduct a *randomized experiment*, as noted by Angrist and Pischke (2009), which coincides with the *mean-independence* assumption in Manski (2003). In that case:

$$\mathbb{E}\left[Y_{ijt}^{(k)} \mid \mathbf{X}_{ijt}, Z_{ijt}=1\right] = \mathbb{E}\left[Y_{ijt}^{(k)} \mid \mathbf{X}_{ijt}, Z_{ijt}=0\right], \quad \text{for } k=0,1. \tag{10}$$

Then, $\delta(\mathbf{X}) = \rho(\mathbf{X})$, and the expression in (7) simplifies to:

$$\Re(\mathbf{X}) = \mathbb{E}[Y_{ijt}^{(1)} \mid \mathbf{X}_{ijt}, Z_{ijt}=1] - \mathbb{E}[Y_{ijt}^{(0)} \mid \mathbf{X}_{ijt}, Z_{ijt}=0]. \tag{11}$$

Hence, under random assignment, $\delta(\mathbf{X})$ suffers no selection bias, and $\Re(\mathbf{X})$ is point-identified.

The mean-independence assumption, however, is rather strict. Suppose now that diversity outcomes are known to satisfy:

$$-\infty < L^{(k)} < Q_Y^{(k)}(p) \leq Y_{ijt}^{(k)} \leq Q_Y^{(k)}(p^c) < U^{(k)} < +\infty, \tag{12}$$

where $Q_Y(p) = \inf\{y : F_Y(y) \geq p\}$, with $F_Y(\cdot)$ the CDF of Y and p^c is the complement of p, i.e., $p^c = 1-p$. Suppose further that we are interested in the average treatment effect rather than the effect for each unit. Using the law of iterated expectations:

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\left[\Re(\mathbf{X})\right] &= \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}\left[Y_{ijt}^{(1)} \mid \mathbf{X}_{ijt}\right]\right] - \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}\left[Y_{ijt}^{(0)} \mid \mathbf{X}_{ijt}\right]\right] \\ &= \mathbb{E}\left[Y_{ijt}^{(1)} - Y_{ijt}^{(0)}\right] = \Re.\end{aligned} \tag{13}$$

Now, since some conditional expectations remain latent (as shown in (8)), one may bound them using either known outcome supports $[L^{(k)}, U^{(k)}]$ or their quantile-based versions $[Q_Y(p), Q_Y(1-p)]$. Manski (1990, 2003) propose the following nonparametric bounds for the treatment effect:

$$\begin{aligned}\Re \in \Big[&\mathbb{E}[Y_{ijt}^{(1)} \mid Z_{ijt}=1]\Pr(Z_{ijt}=1) + L^{(1)}\Pr(Z_{ijt}=0) \\ &- U^{(0)}\Pr(Z_{ijt}=1) - \mathbb{E}[Y_{ijt}^{(0)} \mid Z_{ijt}=0]\Pr(Z_{ijt}=0), \\ &\mathbb{E}[Y_{ijt}^{(1)} \mid Z_{ijt}=1]\Pr(Z_{ijt}=1) + U^{(1)}\Pr(Z_{ijt}=0) \\ &- L^{(0)}\Pr(Z_{ijt}=1) - \mathbb{E}[Y_{ijt}^{(0)} \mid Z_{ijt}=0]\Pr(Z_{ijt}=0)\Big]\end{aligned} \tag{14}$$

These bounds can be tightened by substituting $L^{(k)}$ and $U^{(k)}$ with the quantiles $Q_Y(p)$ and $Q_Y(1-p)$, respectively. In essence, using a similar notation to Manski (2003), the region $H[\Re]$ is the *identification region* for $\Re$, where $H[\Re]$ is defined as the bound in (14). Note that $H[\Re]$ is only partially identified when $0 < \Pr[Z_{ijt} = k] < 1$ for $k = 0, 1$, as otherwise, $H[\Re]$ is simply a singleton. In other words, if, say, $\Pr[Z_{ijt} = 1] = 1$, then both upper and lower bounds coincide with the treated mean, so $H[\Re]$ collapses.

In the following sections, we outline estimation procedures for both the naïve unconditional difference and the nonparametric bounds. We also construct $(1 - \alpha)\%$ confidence intervals for the bounds using Bonferroni-adjusted intervals as proposed by Horowitz and Manski (1998), and derive standard errors via the delta method [see Casella and Berger (2024)].

2.2 Interpretation of the Bounding Constants

The bounding constants

$$L^{(1)} \leq \mathbb{E}\left[Y_{ijt}^{(1)} \mid Z_{ijt} = 0\right] \leq U^{(1)}, \quad L^{(0)} \leq \mathbb{E}\left[Y_{ijt}^{(0)} \mid Z_{ijt} = 1\right] \leq U^{(0)},$$

state that the latent (never-observed) mean outcome a "treated" firm would have realised had it not been treated cannot be lower than $L^{(1)}$ nor higher than $U^{(1)}$; analogously for an "untreated" firm under treatment. Without bounding these counterfactual means the ATE, $\Re$, is not point-identifiable: any value between $-\infty$ and ∞ could be rationalised by suitable (and untestable) choices of $\mathbb{E}[Y_{ijt}^{(1)} \mid Z_{ijt} = 0]$ and $\mathbb{E}[Y_{ijt}^{(0)} \mid Z_{ijt} = 1]$. Because our outcome of interest (Tobin's Q) is unbounded in theory, we adopt quantile-based limits— e.g. $L^{(k)} = Q_Y^{(k)}(0.10)$ and $U^{(k)} = Q_Y^{(k)}(0.90)$— as a reasonable compromise: they confine the worst-case counterfactual means to the central 80% of the empirical outcome distribution, ruling out only the most extreme tail behaviour while introducing minimal additional assumptions. Under these mild restrictions the interval in (14) remains robust to selection on unobservables yet is now finite, so if the entire interval lies above (below) 0 we may still conclude a positive (negative) causal effect even when ignorability fails. We therefore describe $H[\Re]$ as a set of "worst-case bounds" for the ATE under no unverifiable assumptions beyond the outcome range.

2.3 Testing in the Presence of a Random Tipping Point

As is evident from Eq. (1)–(3), the composition of treated and untreated firms depends on the threshold τ. While one could fix τ and analyse the resulting samples, the goal here is different: we seek the *tipping point* at which the average treatment effect $\Re$ becomes strictly positive (or negative). Hence τ must be regarded as a *random* stopping time.

Let $\mathcal{D}_{ijt}$ denote the diversity signal for firm i in sector j at time t (for instance, the percentage of women or non-white executives). A firm is labelled "treated" when $\mathcal{D}_{ijt} \geq \tau$. Rather than prespecify τ, we examine a grid of meaningful cut-offs,

$$\tau_m = m, \quad m \in \mathcal{M}$$

where in our context $\mathcal{M} = \{5, 10, 15, \ldots, 90, 95\}$, with $\overline{\mathcal{M}} = |\mathcal{M}|$ and $m_0 = 5$ and $m_1 = 95$, and, for each m, set

$$Z_{ijt}(m) = \mathbb{1}\{\mathcal{D}_{ijt} \geq \tau_m\}.$$

Thus, we test

$$H_0 : 0 \in H[\Re_u] \quad \forall u \in \mathcal{M} \quad \text{against} \quad H_1 : \exists u \in \mathcal{M} \text{ s.t. } 0 \notin H[\Re_u]. \tag{15}$$

For every τ_m we estimate the corresponding treatment effect using the methods in Sections 3 and 4. The selected threshold $\widehat{\tau}$ is the *smallest* τ_m whose estimated effect differs significantly from zero at the chosen level. Boundary values τ_0 and τ_{100} are excluded, because they would collapse the identification region $H[\Re]$ to a singleton.

Following Siegmund (2013), define the stopping rule

$$\tilde{\tau} = \inf\{\tau : \tau \geq \tau_{m_0}, (H_*[\Re] > 0) \cup (H^*[\Re] < 0)\} \quad (m_0 \geq 5),$$

where $H_*[\Re]$ and $H^*[\Re]$ are the lower and upper bounds of $H[\Re]$. The procedure stops at $\min(\tilde{\tau}, \tau_{m_1})$ and rejects H_0 if $\tilde{\tau} < \tau_{m_1}$ and either $H_*[\Re] > 0$ (positive effect) or $H^*[\Re] < 0$ (negative effect).

Let $S_N := (H_*[\Re] > 0) \cup (H^*[\Re] < 0)$. For a *fixed* threshold τ the test is sized so that $\Pr[S_N \mid H_0] \leq \alpha$. For the random threshold the required family-wise error bound is

$$\begin{aligned} \Pr\Big[&\{(\tilde{\tau} = \tau_{m_0}) \cap S_N\} \bigcup \\ &\{(\tilde{\tau} \in (\tau_{m_0}, \tau_{m_1}]) \cap S_N\} \bigcup \\ &\{(\tilde{\tau} > \tau_{m_1}) \cap S_N\} \Big| H_0 \Big] \leq \alpha \end{aligned} \tag{16}$$

which is equivalent to

$$\Pr\left[\bigcup_{u=m_0}^{m_1} \{(\tilde{\tau} = \tau_u) \cap S_N\} \bigcup \{(\tilde{\tau} > \tau_{m_1}) \cap S_N\} \,\middle|\, H_0 \right] \leq \alpha, \tag{17}$$

and, by Boole's inequality,

$$\begin{aligned} \Pr\left[\bigcup_{u=m_0}^{m_1} \{(\tilde{\tau} = \tau_u) \cap S_N\} \bigcup \{(\tilde{\tau} > \tau_{m_1}) \cap S_N\} \,\middle|\, H_0 \right] \leq \\ \sum_{u=m_0}^{m_1} \Pr\left[(\tilde{\tau} = \tau_u) \cap S_N \mid H_0\right] + \Pr[(\tilde{\tau} > \tau_{m_1}) \cap S_N \mid H_0] \leq \alpha, \end{aligned} \tag{18}$$

so that $\sum_{u=m_0}^{m_1} \alpha_u + \alpha_{m_1+} \leq \alpha$.

Because the grid stops at its largest value τ_{m_1}, the event $(\tilde{\tau} > \tau_{m_1}) \cap S_N$ has probability zero, so the "tail" error share α_{m_1+} can be set to 0 and omitted.

Since all $\overline{\mathcal{M}} = 19$ looks are pre–scheduled and use the same sample, we adopt the equal–spending rule of Pocock (1977):

$$\alpha_u = \frac{\alpha}{\overline{\mathcal{M}}}, \qquad u = m_0, \ldots, m_1, \quad \text{so} \sum_{u=m_0}^{m_1} \alpha_u = \alpha.$$

The two–sided stage-wise critical value is therefore $c_u = \Phi^{-1}(1 - \alpha_u/2) \simeq 3.07$ when $\alpha = 0.05$ and $\overline{\mathcal{M}} = 19$. (Alternative allocations include O'Brien–Fleming (O'Brien and Fleming, 1979) or the Lan–DeMets spending function (Gordon Lan and DeMets, 1983).)

3 Estimation and Identification

In Section 2 we defined the unconditional mean-comparison parameter δ (Eq. (13)) and Manski's bounds $\Re$ (Eq. (14)). We now give their sample analogues and show how to tighten the bounds via quantiles.

The estimator of δ can be written as a single weighted sum:

$$\hat{\delta} = \sum_{i=1}^{n^j} \sum_{j=1}^{K} \sum_{t=1}^{T} Y_{ijt}\, w_{ijt}, \qquad w_{ijt} = \frac{Z_{ijt}}{N_1} - \frac{1 - Z_{ijt}}{N_0},$$

where

$$N_k = \sum_{i,j,t} \mathbb{1}\{Z_{ijt} = k\}, \quad k \in \{0, 1\}, \quad N = N_0 + N_1.$$

since the Central Limit Theorem (CLT hereafter) tells us

$$\sqrt{N}(\hat{\delta} - \delta) \xrightarrow{D} N(0, \sigma^2). \tag{19}$$

To estimate $\Re$, recall from (14) that $\Re$ involves the four quantities $\mathbb{E}[Y^{(k)} \mid Z = k]$ and $\Pr(Z = k)$, $k = 0, 1$, plus the endpoints $\{L^{(k)}, U^{(k)}\}$. We estimate them by

$$\hat{\delta}_k = \frac{1}{N_k} \sum_{i,j,t} Y_{ijt}\, \mathbb{1}\{Z_{ijt} = k\}, \qquad \hat{p}_k = \frac{N_k}{N},$$

and

$$L^{(k)} = \min\{Y_{ijt} : Z_{ijt} = k\}, \quad U^{(k)} = \max\{Y_{ijt} : Z_{ijt} = k\},$$

noting that $\hat{\delta}_1 - \hat{\delta}_0 = \hat{\delta}$ and $\hat{p}_1 = 1 - \hat{p}_0$.

To tighten the raw-support bounds, replace $L^{(k)}, U^{(k)}$ by the sample p- and $(1 - p)$-quantiles in each group:

$$\hat{F}^{(k)}(y) = \frac{1}{N_k} \sum_{i,j,t} \mathbb{1}\{Y_{ijt} \le y,\ Z_{ijt} = k\}, \qquad \hat{Q}_Y^{(k)}(p) = \inf\{y : \hat{F}^{(k)}(y) \ge p\},$$

or equivalently $\hat{Q}_Y^{(k)}(p) = Y_{(\lceil pN_k \rceil)}^{(k)}$ when $Y_{(1)}^{(k)} \le \cdots \le Y_{(N_k)}^{(k)}$ are group-k order stats. Finally, in (14) substitute

$$L^{(k)} \mapsto \hat{Q}_Y^{(k)}(p), \qquad U^{(k)} \mapsto \hat{Q}_Y^{(k)}(1 - p),$$

to obtain the quantile-based bounds.

In Section 4 below we describe how to construct $(1-\alpha)\%$ confidence bands for $\hat{\delta}$ and for the nonparametric bounds via the Bonferroni-adjusted delta-method.

4 Inference

For $u = m_0, \cdots, m_1$, obtaining the $(1-\alpha_u)\%$ confidence intervals for the naïve estimator $\hat{\delta}$ is rather straightforward, since $\hat{\delta}$ is a linear statistic and from (19) it follows that:

$$\sigma^2(\hat{\delta}) = \mathbb{E}\left[(Y_{ijt}w_{ijt} - \delta)^2\right], \tag{20}$$

where

$$\hat{\sigma}^2(\hat{\delta}) = \frac{1}{N-1}\sum_{i,j,t}\left(Y_{ijt}w_{ijt} - \hat{\delta}\right)^2. \tag{21}$$

A $(1-\alpha_u)\%$ Wald-type interval is then

$$\hat{\delta} \pm \Phi^{-1}(1-\alpha_u/2)\sqrt{\hat{\sigma}^2(\hat{\delta})}, \quad \text{for} \quad u = m_0, \cdots, m_1,$$

where $\Phi^{-1}(.)$ is the inverse of CDF of the standard normal distribution.

Obtaining confidence interval for the nonparametric bounds is less straightforward, since the upper/lower estimators are nonlinear. As such we have the following proposition which is predicated on Bonferroni-adjusted delta-method.

Proposition 1. *Let us denote $\mathcal{L}(\hat{\theta})$ and $\mathcal{U}(\hat{\theta})$ as the lower and upper bound estimates of the nonparametric bounds respectively, where $\hat{\theta} = (\hat{\delta}_1, \hat{\delta}_0, \hat{p}_1, \hat{p}_0)^\top$ is a 4×1 vector of estimators. The $(1-\alpha_u)\%$ confidence interval for the union of the bounds is obtained by:*

$$\mathcal{L}(\hat{\theta}) \pm \Phi^{-1}(1-\alpha_u/2) \times S.E.\left(\mathcal{L}(\hat{\theta})\right) \quad \text{and} \quad \mathcal{U}(\hat{\theta}) \pm \Phi^{-1}(1-\alpha_u/2) \times S.E.\left(\mathcal{U}(\hat{\theta})\right) \tag{22}$$

where $Var\left(\mathcal{L}(\hat{\theta})\right) \approx \nabla\mathcal{L}(\theta)^\top \frac{\Omega_{\hat{\theta}}}{N} \nabla\mathcal{L}(\theta)$ *with*

$$\nabla\mathcal{L}(\theta) = (p_1, -p_0, \delta_1 - U_0, L_1 - \delta_0)^\top \tag{23}$$

$$\nabla\mathcal{U}(\theta) = (p_1, -p_0, \delta_1 - L_0, U_1 - \delta_0)^\top \tag{24}$$

and the covariance of the estimators $\hat{\theta}$ is given explicitly by:

$$\Omega_{\hat{\theta}} = \begin{pmatrix} Var(\hat{\delta}_1) & 0 & 0 & 0 \\ 0 & Var(\hat{\delta}_0) & 0 & 0 \\ 0 & 0 & Var(\hat{p}_1) & -Var(\hat{p}_1) \\ 0 & 0 & -Var(\hat{p}_1) & Var(\hat{p}_0) \end{pmatrix}, \tag{25}$$

4.1 Concentration-Driven Confidence Bands for Average Treatment Effects

A major shortcoming of the nonparametric bounds proposed by Manski (1990, 2003) and introduced in Section 2 is the strong assumption that the latent conditional expectations in Eq. (9) lie inside a known bounded interval

$$\left[L^{(k)}, U^{(k)}\right], \quad k \in \{0, 1\}.$$

In practice, these expectations may be unbounded. To address this, we reformulate the problem probabilistically and study

$$\Pr\left(\forall u \in \mathcal{M}, \Re_u \in H_{\alpha_u}[\Re_u]\right).$$

We first analyze the stylized case in which the observations Y_{ijt} are independent for all indices i, j, t. Although independence is useful for pure cross-sectional snapshots, where firms may be assumed to be independent units—for example, analyzing one quarter across many sectors—it is clearly unrealistic in panel settings, where serial correlation may be present. Consequently, we extend our results to the more realistic scenario in which each firm-level time series $(Y_{ij1}, \cdots, Y_{ijT})$ exhibits temporal dependence. In essence, we allow temporal dependence within firm, but assume firms are independent cross-sectionally.

Simplifying the notations in section 2, identification follows Manski (1990, 2003):

$$\Re \in \left[\delta_1 p_1 + L^{(1)} p_0 - U^{(0)} p_1 - \delta_0 p_0,\ \delta_1 p_1 + U^{(1)} p_0 - L^{(0)} p_1 - \delta_0 p_0\right], \tag{26}$$

where $\delta_k := \mathbb{E}[Y_{ijt}^{(k)} \mid Z_{ijt} = k]$ and $p_k := \Pr(Z_{ijt} = k)$. Replacing the latent terms $(\delta_{10}, \delta_{01})$ (where $\delta_{kk'} := \mathbb{E}[Y_{ijt}^{(k)} \mid Z_{ijt} = k']$) by support limits $(L^{(k)}, U^{(k)})$ yields (26). In the first setting, we assume that the data exhibits no cross-sectional or serial dependence.

Assumption 1 (Independent sampling)**.** *The collection $(Y_{ijt}, Z_{ijt})_{i,j,t}$ consists of i.i.d. draws from a sub-exponential distribution.*

Proposition 2 (Finite–sample coverage under i.i.d. sampling)**.** *Let $0 < \alpha_u < 1$ for $u = m_0, \cdots, m_1$ and denote by $N_k = \sum_i \mathbb{1}\{Z_i = k\}$ the sample size in treatment group $k \in \{0, 1\}$ and by*

$$Y_{(1)}^{(k)} \leq \cdots \leq Y_{(N_k)}^{(k)}$$

the order statistics of the observed outcomes in that group. Set

$$\varepsilon_k := \sqrt{\frac{\log(12/\alpha_u)}{2N_k}}, \quad L_{\alpha_u}^{(k)} := Y_{(1)}^{(k)} - \varepsilon_k, \quad U_{\alpha_u}^{(k)} := Y_{(N_k)}^{(k)} + \varepsilon_k, \qquad k = 0, 1.$$

and define the two thresholds

$$t_{p,k} := \sqrt{\frac{\log(12/\alpha_u)}{2N}}, \qquad t_{\mu,k} := \min\left\{ M_k \sqrt{\frac{\log(12/\alpha_u)}{cN_k}}, \frac{M_k}{cN_k} \log\left(\frac{12}{\alpha_u}\right) \right\}.$$

Let $\hat{\mu}_k = \frac{1}{N_k} \sum_{Z_i = k} Y_i$ and $\hat{p}_k = \frac{N_k}{N_0 + N_1}$ be the sample means and treatment shares. Define the random interval

$$H_{\alpha_u}[\Re_u] = \left[\hat{\mu}_1^- \hat{p}_1^- + L_{\alpha_u}^{(1)} \hat{p}_0^- - U_{\alpha_u}^{(0)} \hat{p}_1^+ - \hat{\mu}_0^+ \hat{p}_0^+,\ \hat{\mu}_1^+ \hat{p}_1^+ + U_{\alpha_u}^{(1)} \hat{p}_0^+ - L_{\alpha_u}^{(0)} \hat{p}_1^- - \hat{\mu}_0^- \hat{p}_0^-\right], \tag{27}$$

where $\hat{\mu}_k^{\pm} = \hat{\mu}_k \pm t_{\mu,k}$ and $\hat{p}_k^{\pm} = \hat{p}_k \pm t_{p,k}$ Then, under assumption 1

$$\Pr\left(\forall u \in \mathcal{M}, \Re_u \in H_{\alpha_u}[\Re_u]\right) \geq 1 - \alpha, \tag{28}$$

Proposition 2 states that the data–driven set $H_{\alpha_u}[\Re_u]$ in (29) is a $100(1-\alpha_u)\%$-level confidence region for the average treatment effect $\Re_u$ under nothing more than i.i.d. sampling. Because $\mathbb{E}[Y_{ijt}^{(1)} \mid Z_{ijt} = 0]$ and $\mathbb{E}[Y_{ijt}^{(0)} \mid Z_{ijt} = 1]$ are latent, point

identification is impossible without additional assumptions; the proposition nevertheless guarantees that the random interval constructed from the empirical means, treatment proportions, and slightly "padded" sample extrema will cover the true $\Re_u$ in at least $100(1-\alpha_u)\%$ of repeated samples. Practically, one computes $H_\alpha[\Re]$ by (i) splitting the sample into treated and untreated subsamples, (ii) forming the subsample–specific means μ_k and proportions $\hat{p}_k$, (iii) widening the minimal and maximal observed outcomes by the DKWM envelope $\varepsilon_k = \sqrt{\log(12/\alpha_u)/2N_k}$, and (iv) plugging these objects into (29). The resulting band can be used exactly like an ordinary confidence interval: the null hypothesis $H_0 : \Re_u = r_0$ is rejected at level α_u, whenever $r_0 \notin H_{\alpha_u}[\Re_u]$.

If substantive knowledge implies that the latent outcomes are truncated on one or both tails—for instance, Tobin's Q is bounded below by 0—the extreme-value inputs in Manski's bounds can be replaced by the true population limits. Let λ (lower) and Λ (upper) denote any such known bounds. When both limits are known one sets $a = \lambda$ and $b = \Lambda$ in the plug-in formulas; the resulting $100(1-\alpha)\%$ simultaneous band coincides with Proposition 1 and requires no DKW padding. When only one tail is known—say $Y \geq \lambda$, but the upper support is unknown—we fix the lower extreme at λ while retaining the sample maximum, the DKW envelope on the upper side. The next Corollary shows that this hybrid construction preserves the nominal family-wise coverage probability even when the first significant threshold is data-selected.

Corollary 1 (Finite–sample coverage under i.i.d. sampling and truncated distribution)**.** *Let $0 < \alpha_u < 1$ for $u = m_0, \cdots, m_1$ and denote by $N_k = \sum_i \mathbb{1}\{Z_i = k\}$ the sample size in treatment group $k \in \{0,1\}$ and by*

$$\lambda \leq Y_{(1)}^{(k)} \leq \cdots \leq Y_{(N_k)}^{(k)}$$

the order statistics of the observed outcomes in that group. Set

$$\varepsilon_k := \sqrt{\frac{\log(6/\alpha_u)}{2N_k}}, \quad L^{(k)} := \lambda, \quad U_{\alpha_u}^{(k)} := Y_{(N_k)}^{(k)} + \varepsilon_k,$$

for $k = 0, 1$, and define the two thresholds

$$t_{p,k} := \sqrt{\frac{\log(12/\alpha_u)}{2N}}, \qquad t_{\mu,k} := \min\left\{ M_k \sqrt{\frac{\log(12/\alpha_u)}{cN_k}}, \frac{M_k}{cN_k} \log\left(\frac{12}{\alpha_u}\right) \right\}.$$

Let $\hat{\mu}_k = \frac{1}{N_k} \sum_{Z_i=k} Y_i$ and $\hat{p}_k = \frac{N_k}{N_0+N_1}$ be the sample means and treatment shares. Define the random interval

$$H_{\alpha_u}[\Re_u] = \left[\hat{\mu}_1^- \hat{p}_1^- + L^{(1)} \hat{p}_0^- - U_{\alpha_u}^{(0)} \hat{p}_1^+ - \hat{\mu}_0^+ \hat{p}_0^+,\ \hat{\mu}_1^+ \hat{p}_1^+ + U_{\alpha_u}^{(1)} \hat{p}_0^+ - L^{(0)} \hat{p}_1^- - \hat{\mu}_0^- \hat{p}_0^- \right], \tag{29}$$

where $\hat{\mu}_k^\pm = \hat{\mu}_k \pm t_{\mu,k}$ and $\hat{p}_k^\pm = \hat{p}_k \pm t_{p,k}$ Then, under assumption 1

$$\Pr\left(\forall u \in \mathcal{M}, \Re_u \in H_{\alpha_u}[\Re_u]\right) \geq 1 - \alpha, \tag{30}$$

From here on, we weaken the i.i.d. assumption and allow the data to exhibit weak dependence by assuming each series is a stationary α-mixing process. For example, any stationary AR(1) model satisfies this condition.

Assumption 2 (α-mixing sampling)**.** *The collection $(Y_{ijt}, Z_{ijt})_{i,j,t}$ is a strictly stationary α-mixing process in the sense of Definition 1, with mixing coefficients*

$$\alpha(k) = \sup_{m \geq 1} \alpha\left(\mathcal{B}_1^m,\, \mathcal{B}_{m+k}^{nT}\right),$$

satisfying $\alpha(k) \to 0$ as $k \to \infty$ and $C_\alpha = \sum_{k=1}^{\infty} \alpha(k)^{1/2} < \infty$. Moreover, each outcome Y_{ijt} has a uniformly bounded sub-exponential norm, $\sup_{i,j,t} \|Y_{ijt}\|_{\psi_1} < \infty$.

Proposition 3 (Finite–sample coverage under α–mixing sampling)**.** *Let $0 < \alpha_u < 1$ for $u = m_0, \cdots, m_1$ and let $(Y_{ijt}, Z_{jt})_{i,j,t}$ be a strictly stationary sequence with $Z_i \in \{0, 1\}$, $Y_i \in \mathbb{R}$, and strong–mixing coefficients $\alpha(r)$ satisfying*

$$C_\alpha = \sum_{r=1}^{\infty} \alpha(r)^{1/2} < \infty.$$

Write

$$N_k = \sum_{i=1}^{n} \mathbb{1}\{Z_i = k\}, \quad \hat{p}_k = \frac{N_k}{N_0 + N_1}, \quad \hat{\mu}_k = \frac{1}{N_k} \sum_{Z_i = k} Y_i, \quad k = 0, 1.$$

Define the two thresholds

$$t_{p,k} = (1 + 4C_\alpha)\sqrt{\frac{2\log(12/\alpha_u)}{N_k}}, \qquad t_{\mu,k} = \max\left\{t_k^{(1)}, t_k^{(2)}, t_k^{(3)}\right\},$$

where the $t_k^{(j)}$ are the unique solutions making each term of Lemma 4 bounded by $\alpha_u/18$. Finally, set

$$\varepsilon_k = t_{p,k}, \quad L_{\alpha_u}^{(k)} = Y_{(1)}^{(k)} - \varepsilon_k, \quad U_{\alpha_u}^{(k)} = Y_{(N_k)}^{(k)} + \varepsilon_k, \quad k = 0, 1.$$

Then the random interval

$$H_{\alpha_u}[\Re_u] = \left[\hat{\mu}_1^- \hat{p}_1^- + L_\alpha^{(1)} \hat{p}_0^- - U_\alpha^{(0)} \hat{p}_1^+ - \hat{\mu}_0^+ \hat{p}_0^+,\ \hat{\mu}_1^+ \hat{p}_1^+ + U_\alpha^{(1)} \hat{p}_0^+ - L_\alpha^{(0)} \hat{p}_1^- - \hat{\mu}_0^- \hat{p}_0^-\right]$$

where $\hat{\mu}_k^\pm = \hat{\mu}_k \pm t_{\mu,k}$ and $\hat{p}_k^\pm = \hat{p}_k \pm t_{p,k}$ Then, under assumption 2

$$\Pr\left(\forall u \in \mathcal{M}, \Re_u \in H_{\alpha_u}[\Re_u]\right) \geq 1 - \alpha.$$

Similar to Corollary 1, the extension of Proposition 3 to the case of one-tail truncated latent conditional expectations simply requires modifying the paddings to the bounds. Specifically, for the case of lower tail truncation, replace

$$L_{\alpha_u}^{(k)} := Y_{(1)}^{(k)} - \varepsilon_k \quad \text{and} \quad U_{\alpha_u}^{(k)} := Y_{(N_k)}^{(k)} + \varepsilon_k$$

with

$$L^{(k)} := \lambda, \quad \text{and} \quad U_{\alpha_u}^{(k)} := Y_{(N_k)}^{(k)} + \tilde{\varepsilon}_k,$$

where now

$$\tilde{\varepsilon}_k := (1 + 4C_\alpha)\sqrt{\frac{2\log\left(6/\alpha_u\right)}{N_k}}$$

for $k = 0, 1$, where the observed order statistics satisfy $\lambda \leq Y_{(1)}^{(k)} \leq \cdots \leq Y_{(N_k)}^{(k)}$.

A drawback of the finite-sample bands in Propositions 2 and 3 is that the Bernstein and Hoeffding-type paddings for sub-exponential tails depend on multiple nuisance constants (mixing rates, sub-exponential parameters, etc.), which quickly becomes cumbersome in practice. Moreover, although the sub-exponential assumption is fairly general, it is still a substantive restriction on the data. In Proposition 4 we therefore introduce a hybrid confidence band that combines

- The Dvoretzky–Kiefer–Wolfowitz concentration bound (which requires no tail assumptions beyond finiteness) for the order–statistic endpoints, and
- The usual asymptotic delta-method (CLT) for the sample means and proportions. The DKW inequality controls the uniform deviation $\sup_x |F_n(x) - F(x)|$ in finite samples without any distributional assumptions on Y (see, e.g., Chapter 3 of Van Der Vaart et al. (1996)). This hybrid approach preserves the simplicity of the DKW envelope for the nonparametric piece while relying on asymptotic normality only for the low-dimensional parameters.

Proposition 4 ($100(1-\alpha)\%$ hybrid confidence band under $\boldsymbol{\alpha}$–mixing)**.** *Let $(Y_{ijt}, Z_{ijt})_{i,j,t}$ be strictly stationary with $\boldsymbol{\alpha}$–mixing coefficients $\boldsymbol{\alpha}(r)$ such that*

$$C_\alpha = \sum_{r=1}^{\infty} \boldsymbol{\alpha}(r)^{1/2} < \infty.$$

For

$$Y_{(1)}^{(k)} \leq \cdots \leq Y_{(N_k)}^{(k)},$$

define

$$\varepsilon_k = (1 + 4C_\alpha)\sqrt{\frac{2\log(8/\alpha_u)}{N_k}}, \qquad L_{\alpha_u}^{(k)} = Y_{(1)}^{(k)} - \varepsilon_k, \quad U_{\alpha_u}^{(k)} = Y_{(N_k)}^{(k)} + \varepsilon_k.$$

for $u = m_0, \cdots, m_1$. Let us denote $\mathcal{L}_{\alpha_u}(\hat{\theta})$ and $\mathcal{U}_{\alpha_u}(\hat{\theta})$ as the lower and upper bound estimates of the nonparametric bounds respectively, where $\hat{\theta} = (\hat{\delta}_1, \hat{\delta}_0, \hat{p}_1, \hat{p}_0)^\top$ is a 4×1 vector of estimators. The $100(1-\alpha)\%$ confidence interval for the union of the bounds is obtained by:

$$\mathcal{L}_{\alpha_u}(\hat{\theta}) \pm \Phi^{-1}(1-\alpha_u/4).S.E.\left(\mathcal{L}_{\alpha_u}(\hat{\theta})\right) \quad \text{and} \quad \mathcal{U}_{\alpha_u}(\hat{\theta}) \pm \Phi^{-1}(1-\alpha_u/4)S.E.\left(\mathcal{U}_{\alpha_u}(\hat{\theta})\right) \tag{31}$$

where $Var\left(\mathcal{L}(\hat{\theta})\right) \approx \nabla\mathcal{L}(\theta)^\top \frac{\Omega_{\hat{\theta}}}{N} \nabla\mathcal{L}(\theta)$ with

$$\nabla\mathcal{L}(\theta) = \left(p_1, -p_0, \delta_1 - U_{\alpha_u}^{(0)}, L_{\alpha_u}^{(1)} - \delta_0\right)^\top \tag{32}$$

$$\nabla\mathcal{U}(\theta) = \left(p_1, -p_0, \delta_1 - L_{\alpha_u}^{(0)}, U_{\alpha_u}^{(1)} - \delta_0\right)^\top \tag{33}$$

and the covariance of the estimators $\hat{\theta}$ is given explicitly by:

$$\Omega_{\hat{\theta}} = \begin{pmatrix} Var(\hat{\delta}_1) & 0 & 0 & 0 \\ 0 & Var(\hat{\delta}_0) & 0 & 0 \\ 0 & 0 & Var(\hat{p}_1) & -Var(\hat{p}_1) \\ 0 & 0 & -Var(\hat{p}_1) & Var(\hat{p}_0) \end{pmatrix}, \tag{34}$$

Following the same logic as Corollary 1, the extension of Proposition 4 to the case of one-tail truncated latent conditional expectations entails replacing

$$L_{\alpha_u}^{(k)} := Y_{(1)}^{(k)} - \varepsilon_k \quad \text{and} \quad U_{\alpha_u}^{(k)} := Y_{(N_k)}^{(k)} + \varepsilon_k$$

with

$$L^{(k)} := \lambda, \quad \text{and} \quad U_{\alpha_u}^{(k)} := Y_{(N_k)}^{(k)} + \tilde{\varepsilon}_k,$$

where now

$$\tilde{\varepsilon}_k := (1 + 4C_\alpha)\sqrt{\frac{2\log(4/\alpha_u)}{N_k}},$$

for $k = 0, 1$.

5 Monte Carlo Study

To study the finite-sample behaviour of the hybrid band in Proposition 4 we run a Monte-Carlo experiment with seven data-generating processes (DGPs). Each design is replicated $B = 2{,}000$ times on a single sector with $n = 50$ firms observed for $T \in \{1, 2, 5\}$ periods, giving sample sizes $N = nT \in \{50, 100, 250\}$. A single diversity cut-off $\tau^\circ = 50\%$ is analysed; hence no Bonferroni size split is required. The overall two-sided size is fixed at $\alpha = 0.05$, giving the critical values

$$c_M = \Phi^{-1}(1 - \alpha/2) \approx 1.96 \quad \text{and} \quad c_H = \Phi^{-1}(1 - \alpha/4) \approx 2.24$$

correspondingly for the Manski and Hybrid approaches. The realised outcome is

$$Y_{it}^{\text{obs}} = Y_{it}^0 + \Delta D_{it}, \qquad \Delta = 4,$$

where Y_{it}^0 follows the distribution listed below and $D_{it} \sim \text{Bernoulli}(0.3)$.

DGP A: *i.i.d. Standard Normal*
$Y_{it}^0 \sim N(0, 1)$, $D_{it} \sim \text{Bernoulli}(0.3)$.

DGP B: *Heavy tail (sub-exponential)*
$Y_{it}^0 \sim t_3/\sqrt{3}$ (unit variance), $D_{it} \sim \text{Bernoulli}(0.3)$.

DGP C: *AR(1) panel with **negative** selection bias*
$Y_{it}^0 = 0.4Y_{i,t-1}^0 + \varepsilon_{it}$, $\varepsilon_{it} \overset{\text{i.i.d.}}{\sim} N(0, 1)$.
Treatment probability:
$\Pr(D_{it} = 1 \mid Y_{it}^0) = \text{logit}(-0.5Y_{it}^0 + \eta_{it})$, $\eta_{it} \sim N(0, 0.5^2)$.

DGP D: *AR(1) panel with **positive** selection bias*
$Y_{it}^0 = 0.4Y_{i,t-1}^0 + \varepsilon_{it}$, $\varepsilon_{it} \overset{\text{i.i.d.}}{\sim} N(0, 1)$.
Treatment probability:
$\Pr(D_{it} = 1 \mid Y_{it}^0) = \text{logit}(+0.5Y_{it}^0 + \eta_{it})$, $\eta_{it} \sim N(0, 0.5^2)$.

DGP E: *Rare–extreme point mass (unseen support)*

$$Y_{it}^0 = \begin{cases} -10 & \text{w.p. } 0.002, \\ Z & \text{w.p. } 0.996, \ Z \sim N(0, 1), \\ +10 & \text{w.p. } 0.002, \end{cases} \qquad D_{it} \sim \text{Bernoulli}(0.3).$$

DGP F: *Left-truncated χ^2 tail*

$$Y_{it}^0 \sim \chi^2(3), \qquad D_{it} \sim \text{Bernoulli}(0.3).$$

DGP G: *Uniform support known a priori*
$Y_{it}^0 \sim \text{Uniform}[-5, 5]$, $D_{it} \sim \text{Bernoulli}(0.3)$.
Note: the estimator is supplied with the true support $a = -5$, $b = 5$ when forming Manski bounds.

DGP E is specifically constructed so that the finite sample has a high probability of not observing the population extrema, the precise scenario for which the hybrid band was developed.

5.1 Simulation Results

Table 5.1: Simultaneous coverage of the 95% hybrid and Manski bands

	$N = 50$		$N = 100$		$N = 250$	
DGP	Hybrid	Manski	Hybrid	Manski	Hybrid	Manski
A Standard normal	85.95	9.25	89.75	21.40	96.05	49.50
B t_3 heavy-tail	83.05	36.70	93.40	66.65	99.65	94.90
C AR(1) bias (–)	99.40	51.10	100	84.05	100	99.70
D AR(1) bias (+)	100	84.00	100	98.35	100	100
E Large extrema	89.10	27.30	93.65	46.45	98.80	81.80
F χ^2_3	100	99.85	100	100	100	100
G Uniform	100	100	100	100	100	100

Table 5.1 shows that the hybrid band fulfills its intended role whenever the finite sample is likely to miss the population extrema, while coinciding with the Manski interval in designs where the support is fully known. For the light-tailed Normal benchmark (DGP A) the plug-in Manski interval captures the true effect in barely one tenth of the $B = 2{,}000$ replications for a sample size of $N = 50$ and still under-covers at $N = 100$. Adding the DKW pad and lifts hybrid coverage into the mid-80 percent range for $N = 50$ and drives it close to the nominal 95 percent by $N = 250$. The same pattern holds for the t_3 heavy-tail (DGP B): Manski improves as the sample begins to see extreme draws, but hybrid is uniformly closer to the target and reaches virtually perfect coverage in the largest sample.

Serial dependence and endogenous treatment (DGPs C and D) widen both bands. Under negative selection bias, Manski still misses the effect in half of the small-sample replications, whereas hybrid covers more than 99% of the time; with positive selection both bands converge, highlighting that the coverage gap arises specifically when the sample fails to capture the relevant tails. That point is most evident in the rare-extreme design (DGP E): the population includes outcomes of ± 10 with probability only 0.2%, so the finite sample almost never observes them; Manski therefore covers only 27% at $N = 50$, whereas the hybrid correction restores coverage to 89% and rises above 98% by $N = 250$.

When the lower support is known to be zero, as under the left-truncated $\chi^2(3)$ baseline (DGP F), only the upper-tail pad is required and Manski already attains nominal coverage; hybrid is effectively the same band. The same coincidence is observed for the uniform distribution with fully known support (DGP G), where both methods hit 100 percent in every cell. Taken together, the results corroborate the theory: hybrid bands deliver the promised finite-sample protection precisely in situations where the classical plug-in Manski interval is too narrow and reduce to Manski when no tail uncertainty remains.

6 Empirical Application

In this section, we ask "Does gender-based board diversity causally affect firm innovation?". We begin by outlining the data and summarizing its key descriptive statistics. We then present the nonparametric bounds approach of Manski (1990, 2003) with simultaneous confidence bands in Proposition 1, the hybrid band proposed in Proposition 4, and the naïve mean-comparison framework of Angrist and Pischke (2009) (reported in Appendix D). The common objective is to test the null hypothesis of a zero average treatment effect of diversity on innovation—against a positive or negative alternative—when the diversity cut-off is selected endogenously (see Eq. (15)).

6.1 Data and Descriptive Statistics

The empirical analysis uses a panel of publicly listed firms compiled from FactSet, with quarterly observations from 2015 Q2 through 2022 Q1. The initial sample includes 945 firms, yielding a short panel of 945 cross-sectional units over 28 quarters (totalling 26,460 firm-quarter observations).

In our analysis, we categorize the eleven GICS sectors into five broader groups: Cyclicals (Consumer Discretionary, Materials, Industrials, Real Estate), Defensives (Health Care, Consumer Staples, Utilities), Growth & Innovation (Information Technology, Communication Services), Financials, and Energy. This classification reflects the economic sensitivities of these sectors, as identified by MSCI. Specifically, MSCI's Cyclical and Defensive Sectors Indexes classify sectors based on their performance correlation with the business cycle, using the OECD Composite Leading Indicator. According to MSCI, sectors like Consumer Discretionary, Materials, Industrials, Real Estate, Information Technology, Communication Services, and Financials are considered cyclical due to their positive correlation with economic expansions. Conversely, sectors such as Health Care, Consumer Staples, Utilities, and Energy are deemed defensive, exhibiting resilience during economic downturns. By adopting this grouping, we aim to capture the nuanced behaviors of these sectors in relation to macroeconomic conditions, facilitating a more informed analysis of sectoral dynamics. This classification can be found in table D.1.

Following the CSRD definition of a 'large undertaking' (Directive 2022/2464/EU, Art. 3 Pt 4) and the 250-employee threshold used in EU and UK gender-pay-gap statutes, we restrict the sample to firms whose time-average workforce is at least 250 employees over the sample horizon to ensure they fall under harmonized dis-

closure regimes. The restriction yields $n = 901$ firms and a total of $N = 25,228$ firm-quarter observations.

Variable	Min	Mean	Median	Max	Std Dev	Skewness	Kurtosis	N
(%) Women	0.000	27.490	27.140	100	12.723	0.463	1.949	25,038
(%) Unknown gender	0.000	0.029	0.023	0.496	0.031	2.348	14.420	25,038
Tobin's Q	-0.612	0.445	0.012	5.047	1.221	2.133	4.530	23,085
Total assets	10.392	16.109	16.114	22.098	1.811	0.060	0.221	23,990
Leverage	0.000	0.302	0.288	3.945	0.230	3.283	34.292	23,977
Total employees	85.559	25707.813	8554.289	941046.440	54162.070	6.216	58.997	25,224

Table 6.1: Panel descriptive statistics

Note: N varies by variable because some firm–quarter observations are missing that particular item (e.g. Tobin's Q is reported for 23,085 of the 25,228 firm–quarters). The descriptive statistics are computed on all available values for each variable ("pair-wise" basis). For the causal analysis we use listwise deletion, retaining only the firm–quarters for which *all* diversity indicators and Tobin's Q are present.

The key "treatment" variable is the percentage of women in senior leadership positions. These diversity measures are constructed using a supervised machine-learning algorithm applied to senior executives' names, which infers gender from linguistic patterns. If the algorithm cannot assign a gender with high confidence, the individual is labeled as "unknown". Importantly, the incidence of unknown classifications is very low: on average only about 0.03% (Table 6.1). The outcome of interest is Tobin's Q, defined as the ratio of the firm's market value to the replacement cost of its assets, a standard measure of firm performance and growth opportunities (Tobin, 1969, 1978). We also utilize several control variables for descriptive analysis, including firm size (log total assets), leverage (debt-to-assets ratio), and total employees. Summary statistics for all main variables are provided in Table 6.1. After excluding observations with missing data on key fields, the average percentage of women in senior roles is about 27.5%. The standard deviation (around 12 percentage points for female share) indicates considerable cross-firm variation. Notably, a non-trivial subset of firm-quarters have zero diversity: roughly 4% of observations have no women in senior leadership, at least at some point in the sample. The distribution of the diversity variables is right-skewed. Figure 6.1 illustrates kernel density estimates of the percentage of female senior leaders across all firm-quarters. The distribution is skewed to the right with a primary mode around 25–35%, and a secondary mass at 0% corresponding to firms and periods with homogeneous leadership teams.

We next explore the raw association between gender diversity measure and firm performance. In the full sample (pooled across all sectors and time periods), there is a strong positive correlation between senior-team diversity and Tobin's Q. Figure 6.2 plots rolling correlations over time, using a moving window of half the sample period ($T/2 \approx 14$ quarters) to track how the relationship evolves. The Pearson correlation between the percentage of women in leadership and Tobin's Q is in the range of $+0.6$ to $+0.7$ for most of the sample, indicating a fairly strong linear association. The figure also reports Kendall's τ rank correlation, which captures monotonic association; this measure corroborates the positive link while being slightly lower in magnitude, suggesting the relationship is broadly monotonic even if not perfectly linear. The association appears to strengthen from 2015 up to about 2019, consistent with increasing awareness and implementa-

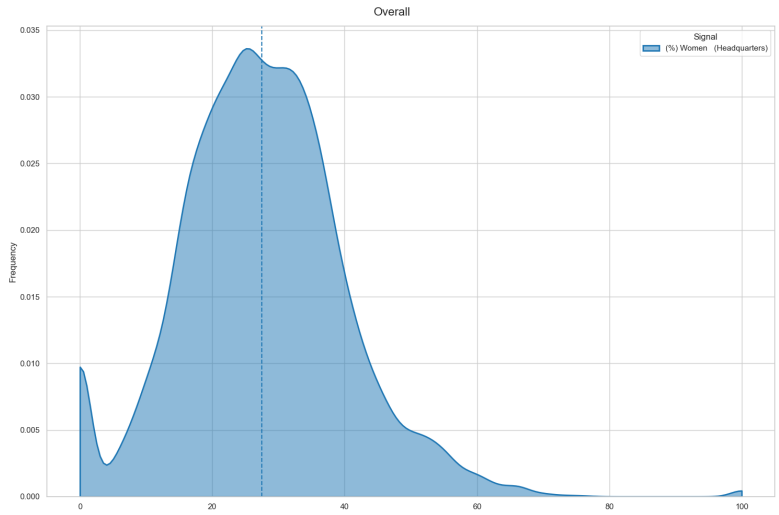


Figure 6.1: Kernel density plot of percentage women. Scott's rule (Scott, 2015) is used to select the smoothing bandwidth parameter.

tion of diversity initiatives, but then shows a noticeable drop around 2019–2020. After 2019, the rolling correlations decline, implying that the previously tight diversity–performance relationship loosened and is increasing again after 2021. One possible interpretation is that external shocks or changing market conditions (for instance, the disruptive impact of the COVID-19 pandemic or the murder of George Floyd) temporarily weakened the correlation between diversity and market valuations.

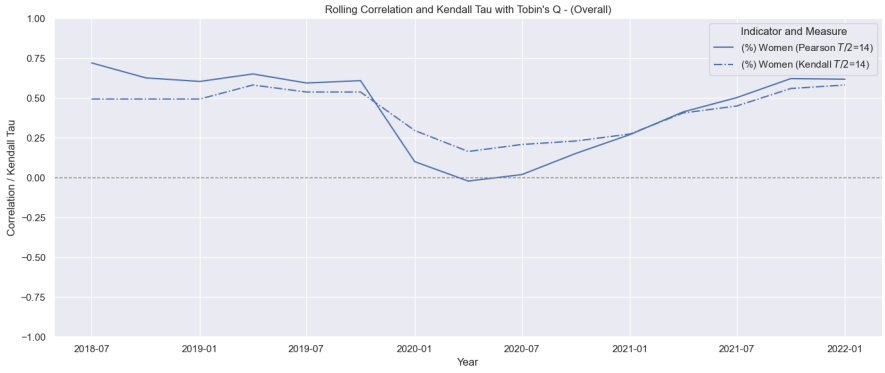


Figure 6.2: Rolling Pearson correlations and Kendall's τ capturing both linear and monotonic associations between Tobin's Q and percentage women in senior leadership.

Note: The size of the rolling windows is chosen as half the length of the time dimension of the sample, i.e., $T/2$.

In light of this, and in addition to the sectoral group analysis, we examine the overall rolling correlations for the period preceding this drop. The corresponding Pearson correlations for this classification are reported in Figure D.2, while the rolling associations are shown in Figure D.3.

Several noteworthy patterns emerge. First, the Growth & Innovation sector consistently exhibits a strong positive correlation between gender diversity measures and Tobin's Q across all years, and this sector does not experience the 2019

drop in correlation seen in the aggregate data. Second, the Energy sector shows a markedly different pattern: the percentage of women in senior positions in energy firms is actually negatively correlated with Tobin's Q in most years. These observations may reflect unique dynamics or reverse causality in the energy industry (for example, struggling firms may appoint more women to leadership roles as part of restructuring). Third, in the Financials sector, the correlation with diversity is negative in the earlier part of the sample (implying more homogenous banks were associated with slightly higher Q ratios pre-2019), but this relationship reverses sign around 2019. By the end of the sample period, financial firms with more diverse leadership tend to have higher Tobin's Q, indicating a possible structural change in how markets value diversity in finance or how an increase in inclusion that enabled diversity to be leveraged for business gains.

6.2 Causality Analysis

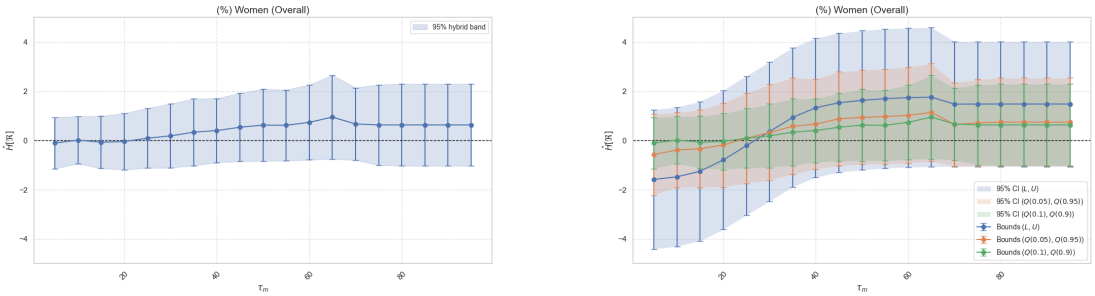
While the descriptive results suggest a concordance between greater senior-level diversity and higher firm performance, correlation alone cannot establish causality. In this section, we formally test whether increases in executive diversity causally impact Tobin's Q, using the methodology developed in Sections 2–4. Because the "treatment" (crossing a diversity threshold) is not randomly assigned, a naïve estimation of this effect risks bias from selection on unobservables. We therefore implement both a conventional point-estimation approach under strong assumptions and a robust partial-identification approach under minimal assumptions, and compare the findings.

First, we apply an unconditional mean-comparison framework following Angrist and Pischke (2009). For each candidate diversity threshold τ (e.g. 5%, 10%, ..., 50%, etc.), firms are split into a treated group (above the threshold) and a control group (below the threshold). We then estimate the difference in mean Tobin's Q between treated and control firms for that threshold. This difference-in-means is a point estimate of the ATE if one assumes mean independence (i.e. that, conditional on crossing the threshold, potential outcomes are the same for treated and control firms on average). We construct simultaneous 95% confidence bands for these ATE estimates across all thresholds in the set $M = 5, 10, 15, \ldots, 90, 95$, applying a Bonferroni or Šidák correction to account for the multiple comparisons. This yields a series of tests for the null hypothesis of no effect at each diversity level, adjusted so that the overall family-wise error rate is 5%. It is important to note that this point-identified approach treats the threshold "treatment" as if random; in practice, firms that surpass a given diversity level could differ systematically from those that do not (for instance, more progressive or better-governed firms might both adopt diverse leadership and perform well for other reasons). As a result, the point estimates of δ may capture more than the true causal effect of diversity. We use this method as a benchmark, fully aware that its validity hinges on strong assumptions.

We next relax the strong assumptions by employing a partial identification strategy (Manski, 1990, 2003). Instead of assuming we can precisely identify the counterfactual outcome for each firm, we derive bounds on the possible ATE. Instead of point identification, we partially identify the region in which the average treatment effect $\Re$ lies, as characterized by Eq.(14). We denote this set

the identification region $H[\Re_u]$ for all u in $\mathcal{M}$, where as noted in Section 2.3, $\mathcal{M} = \{5, 10, 15, \cdots, 90, 95\}$ which represents the random diversity thresholds. As previously noted, estimation of Eq. (14) involves latent quantities $\mathbb{E}[Y_{ijt}^{(0)} \mid Z_{ijt} = 1]$ and $\mathbb{E}[Y_{ijt}^{(1)} \mid Z_{ijt} = 0]$, which are not observed but can be bounded by quantities $L^{(k)}$ and $U^{(k)}$. On on hand, we may acknowledge that the extrema of the latent outcomes within the finite sample may not capture the true population extrema (and consequently the true treatment effect interval), in which case we rely on the finite sample hybrid approach. On the other hand, one may argue that since using the full range of outcomes (min and max) can lead to overly conservative bounds, we also construct Manksi bounds using the (5^{th}, 95^{th}) and (10^{th}, 90^{th}) quantiles of $Y_{ijt}^{(k)}$. Finally, we build a simultaneous joint 95% confidence region for the estimated bounds to make causal inference claims.

Before turning to results, we address some practical implementation details. As noted in Section 3, it is necessary for both the treated and control groups to be non-empty (and sufficiently large) at each threshold to estimate meaningful effects. In our panel, some extreme diversity thresholds (especially very high ones) result in very few firms in one group. We therefore discard threshold levels τ for which one of the groups contains fewer than 10 observations (approximately, we require at least 10 firm-quarters above and below the threshold). If too many high-τ values are discarded for a particular subset of the data, that subset is excluded from the threshold analysis due to lack of support. In practice, this means that for some sector-specific analyses we cannot evaluate very high diversity percentages because, for example, no firm in a given sector ever reaches 90% female leadership. Based on this criterion, certain combinations of sector and diversity type are dropped from the causal analysis. In particular, we exclude female leadership in sectors that never approach gender parity (notably the Financials and Energy sectors). These exclusions are a matter of data availability and ensure that the identification regions for ATE do not trivially collapse to a point. All remaining sector clusters and diversity measures satisfy $0 < \Pr(Z = 1) < 1$ at the thresholds of interest, so both treated and control outcomes can be observed in those cases.



(a) Hybrid: Overall (%) women

(b) Manski: Overall (%) women

Figure 6.3: Hybrid and Manski's Nonparametric Bounds - (Overall)

Note: The lines in the nonparametric bounds plots represent the midpoints between the upper and lower bound estimates.

We first discuss the overall sample (pooling all industries). The naive point estimates suggest that higher diversity leads to better performance, with a tipping point emerging at a moderate level of female representation. In particular,

the Angrist–Pischke mean comparison indicates that once female executives make up roughly one-third of the top team, the estimated ATE on Tobin's Q becomes positive and statistically significant (the 95% confidence band excludes zero). This is evidenced by a significant jump at about the 30–35% female share in leadership (see Table 6.2, which summarizes the estimated threshold levels at which the treatment effect becomes significant under each method). These point-identification results, however, must be interpreted cautiously, since they assume no unobserved confounding. Indeed, the partial identification analysis reveals that under weaker assumptions the evidence is less definitive. The Manski bounds and our hybrid concATE bands for the overall sample show an increasing trend as diversity rises, but crucially the confidence band for the ATE always includes zero at conventional confidence levels. The estimated bounds exhibit a sigmoid-like shape: at very low diversity levels, the lower bound on the ATE is substantially negative (reflecting the possibility that token diversity could harm performance or that the most homogeneous firms might be unusually strong performers), but this lower bound rises toward zero as diversity increases. We observe an inflection point around approximately 20–25% female representation, beyond which the bounds begin to narrow. This value is close to the "critical mass" threshold theorized by Kanter (1977) – the point at which a minority group's representation shifts from tokenism to a more influential presence. Above roughly 25% female share, the worst-case (lower-bound) effect of diversity is no longer hugely negative; it hovers near zero, while the upper-bound effect is positive. Nevertheless, even at the highest levels of diversity observed (e.g. 80–90% female leadership), the 95% joint confidence region for the ATE bounds still straddles zero under the baseline (full-range) Manski scenario. In other words, without stronger assumptions we cannot conclusively assert a positive causal effect in the full sample – the data are consistent with a benefit from diversity, but also with no effect. This underscores the importance of robust inference: what appears significant in the point estimate can become statistically ambiguous once we account for uncertainty about counterfactual outcomes.

If we incorporate mild additional assumptions by trimming the outcome tails, the partial identification results become slightly more optimistic, though still cautious. For example, imposing that the true outcome lies within the 5th and 95th percentile of observed Tobin's Q (thereby excluding implausibly extreme counterfactuals) yields somewhat tighter bounds. In this case we find that at high diversity levels (for instance, beyond 60–70% female leadership) the lower bound of the ATE nearly exceeds zero. Using the 10th and 90th percentile restrictions – a stronger assumption that rules out the extreme 10% tails – we even see the lower bound move just above zero for some thresholds. However, these effects are marginal, and at the 95% confidence level the concATE band for the overall sample still does not fully exclude zero for any threshold when using only the weakest (10th/90th) trimming. The overall conclusion is that, in the full sample, the positive relationship between diversity and performance could be causal, but a conservative analysis cannot rule out a zero impact. The naive method's significant findings (e.g. a female-share tipping point around one-third) may reflect underlying selection bias or favorable unobserved characteristics of diverse firms, since those findings disappear when we allow for more uncertainty.

We also examine the causal effects within certain sector groupings, focusing

Sector	Signal	Hybrid	Manski			Angrist
			Max	5%	10%	
Overall						
	(%) Women	-	-	-	-	35%
Cyclicals						
	(%) Women	-	-	-	-	30%
Defensives						
	(%) Women	-	-	-	-	40%
Growth & Innovation						
	(%) Women	55%	-	-	55%	-
Financials						
	(%) Women			N/A		
Energy						
	(%) Women			N/A		

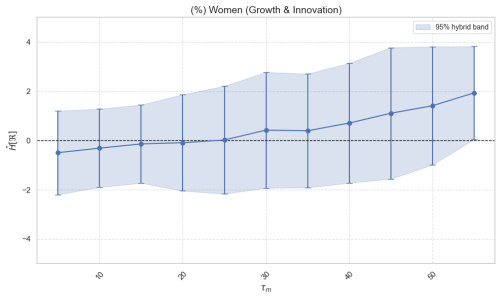
Table 6.2: Random Diversity Tipping Points

Note: This table presents the estimated tipping points—i.e., the random diversity thresholds at which the diversity treatment has a significantly positive effect on Tobin's Q. Cells marked with a (–) indicate cases where significance is not achieved at any of the prescribed thresholds. Rows labeled "N/A" correspond to cases that do not meet the minimum threshold size condition of $\tau_m > 50$ discussed earlier.

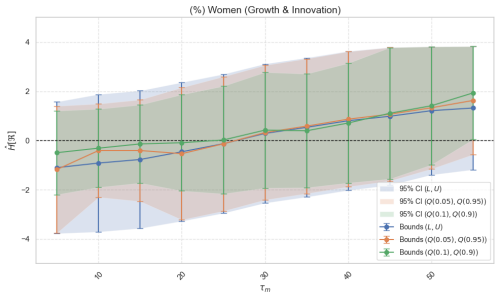
on cases where sufficient diversity variation exists. Notably, the Cyclicals sector (which includes industries like Consumer Discretionary and Industrials) shows a clearer pattern of a diversity tipping point. In this sector, our hybrid partial-identification band indicates that once women comprise more than about 55% of senior management – i.e. when female leadership surpasses men – the ATE on Tobin's Q becomes positive and statistically distinguishable from zero. In other words, for cyclically sensitive industries, achieving majority-female leadership is associated with a robust increase in firm value. Interestingly, the naïve point estimate in cyclicals also eventually signals a positive effect of female leadership, but it finds significance at a somewhat lower threshold (around 30% in our data). This discrepancy highlights how the point estimate can give a premature indication of significance by not accounting for potential biases; the concATE band insists on a higher threshold (and thus a larger performance gain) before declaring the effect significant, reflecting a more stringent standard of evidence. Turning to the Growth & Innovation sector (technology and communications firms), female leadership also appears to have a strong effect once it reaches a critical mass. The hybrid confidence band for the female-share ATE in growth industries becomes significantly positive at roughly 50–55% (about equal gender representation). This aligns with the idea that innovative firms may harness diverse perspectives especially well once a balanced team is in place. The point-estimate analysis also reflects a positive effect in this sector, and interestingly it suggests significance starting at approximately the same range (around the 50% threshold), reinforcing the partial-bounds finding in this case.

In more traditional or constrained sectors such as Defensives and Energy, the lack of sufficient high-diversity observations means our method finds no significant effects. In the Defensive industries (e.g. utilities, healthcare), few firms exceeded

40% female leadership during the sample, and accordingly neither the naive nor the robust approach indicates any clear performance gains from diversity — indeed, the concATE bounds always remain wide and centered around zero. In Energy, as noted, diversity levels are generally low and sometimes inversely related to performance in simple correlations. Consistently, we do not find any positive causal effect in Energy firms; if anything, the point estimates for female leadership in Energy were negative (though not significant under bounds).



(a) Hybrid: Growth & innovation (%) women



(b) Manski: Growth & innovation (%) women

Figure 6.4: Hybrid and Manski's Nonparametric Bounds, and Angrist's Point Estimates - (Significant Results)

Note: The lines in the nonparametric bounds plots represent the midpoints between the upper and lower bound estimates.

Finally, in the Financials sector, female representation never reached high levels so the female diversity effect is not estimable there with our approach (we excluded that case).

In summary, the causality analysis using nonparametric bounds and finite-sample confidence bands paints a more nuanced picture than the raw correlations. The data provide qualified evidence of "tipping points": in certain high-growth or cyclical environments, reaching a critical mass of diversity (for example, women comprising about half of senior leadership roles) is associated with a reliable increase in firm value. However, under the robust inference approach, these effects emerge at higher thresholds and with less ubiquity than a naïve analysis would suggest. The concATE methodology ensures that any detected effect is robust to heavy-tailed outcomes and potential selection bias, guarding against false positives. Where the conventional method finds significance at lower diversity levels, the bounded analysis often still intersects zero, meaning we cannot rule out an absence of effect without further assumptions.

7 Concluding Remarks

This paper introduces concATE as a general framework for robust causal inference when point identification is not possible or reliable. By marrying Manski's (1990; 2003) nonparametric bounds with finite-sample concentration inequalities, concATE offers researchers a new tool to obtain ATE confidence bands without assuming away heavy-tailed outcomes or requiring strong parametric models. The methodology's broader relevance lies in its ability to deliver valid inference under minimal assumptions and even with weakly dependent data, thereby guarding

against false positives that can arise from conventional point estimates under misspecified models or overlooked tail risks.

Our empirical findings on workforce diversity illustrate the importance of such rigorous inference. While naïve regressions might suggest that even modest increases in female leadership yield significant gains, the concATE approach paints a more nuanced picture. We find that substantive benefits of gender diversity materialize only once a sufficient representation level is achieved. In practice, this means token diversity (e.g. a lone female or two in senior leadership) is unlikely to drive measurable performance improvement, whereas reaching a critical mass of women in leadership – roughly one-third or more in growth-oriented industries (and higher in others) – is associated with a reliably positive impact on firm value. These conclusions align with the critical mass hypothesis: diversity can boost performance, but only after crossing a threshold that moves an organization beyond tokenism (Kanter, 1977). By confirming this pattern under stringent inference, our study provides guidance for firms and policymakers – emphasizing that real gains from diversity require either significant numbers of women or alternatively substantial inclusion efforts[1] – and also demonstrates how concATE can be applied in other domains to uncover robust causal insights where traditional methods may be misleading.

References

Adams, R. B. and Ferreira, D. (2009). Women in the boardroom and their impact on governance and performance. *Journal of financial economics*, 94(2):291–309.

Ali, M., Kulik, C. T., and Metz, I. (2011). The gender diversity–performance relationship in services and manufacturing organizations. *The International Journal of Human Resource Management*, 22(07):1464–1485.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Brainard, W. C. and Tobin, J. (1968). Pitfalls in financial model building. *The American economic review*, 58(2):99–122.

Casella, G. and Berger, R. (2024). *Statistical inference*. CRC press.

Dedecker, J. and Merlevède, F. (2007). The empirical distribution function for dependent variables: asymptotic and nonasymptotic results in. *ESAIM: Probability and Statistics*, 11:102–114.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669.

[1]We note that an inclusive organizational culture may allow firms to reap performance gains at lower diversity levels than this critical mass. In settings without an inclusive culture, a small number of women leaders often remain marginalized "tokens" with limited influence, but when genuine inclusion is present, even a few women in leadership can contribute meaningfully to performance improvements. It is an area for future research to investigate the relationship between inclusion and diversity.

Gordon Lan, K. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663.

Hambrick, D. C. and Mason, P. A. (1984). Upper echelons: The organization as a reflection of its top managers. *Academy of management review*, 9(2):193–206.

Hoeffding, W. (1994). Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426.

Hoogendoorn, S., Oosterbeek, H., and Van Praag, M. (2013). The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management science*, 59(7):1514–1528.

Horowitz, J. L. and Manski, C. F. (1998). Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *Journal of Econometrics*, 84(1):37–58.

Kanter, R. M. (1977). Some effects of proportions on group life: Skewed sex ratios and responses to token women. *American journal of Sociology*, 82(5):965–990.

Kanter, R. M. (1987). Men and women of the corporation revisited. *Management Review*, 76(3).

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*, volume 61. Springer.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V: The Luminy Volume*, volume 5 of *Institute of Mathematical Statistics Collections*, pages 273–292. IMS.

Nathan, M. and Lee, N. (2013). Cultural diversity, innovation, and entrepreneurship: firm-level evidence from london. *Economic geography*, 89(4):367–394.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556.

Østergaard, C. R., Timmermans, B., and Kristinsson, K. (2011). Does a different view create something new? the effect of employee diversity on innovation. *Research policy*, 40(3):500–509.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199.

Post, C. and Byron, K. (2015). Women on boards and firm financial performance: A meta-analysis. *Academy of management Journal*, 58(5):1546–1571.

Rio, E. (2000). Inégalités de hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(10):905–908.

Safiullah, M., Akhter, T., Saona, P., and Azad, M. A. K. (2022). Gender diversity on corporate boards, firm performance, and risk-taking: New evidence from spain. *Journal of Behavioral and Experimental Finance*, 35:100721.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Siegmund, D. (2013). *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media.

Tobin, J. (1969). A general equilibrium approach to monetary theory. *Journal of money, credit and banking*, 1(1):15–29.

Tobin, J. (1978). Monetary policies and the economy: the transmission mechanism. *Southern economic journal*, pages 421–431.

Tobin, J. and Brainard, W. C. (1976). Asset markets and the cost of capital.

Torchia, M., Calabrò, A., and Huse, M. (2011). Women directors on corporate boards: From tokenism to critical mass. *Journal of business ethics*, 102:299–317.

Van Der Vaart, A. W., Wellner, J. A., van der Vaart, A. W., and Wellner, J. A. (1996). *Weak convergence*. Springer.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

White, H. (2014). *Asymptotic theory for econometricians*. Academic press.

A Lemmas

In this Section, we introduce the lemmas used in the proofs of the finite-sample propositions and corollaries. The first set of lemmas corresponds to Assumption 1, where the data is assumed to be independent and drawn from a sub-exponential distribution. The second set pertains to Assumption 2, which allows for weakly dependent data.

A.1 Independent Data

In what follows, we introduce the lemmas that provide the concentration inequalities for the estimators and latent quantities involved in the nonparametric bounds. The generalized Bernstein inequality for sub-exponential variables is taken from Vershynin (2018), the Hoeffding bound for Bernoulli random variables from Hoeffding (1994), and the Dvoretzky–Kiefer–Wolfowitz inequality from Kosorok (2008).

Lemma 1 (Bernstein inequality for i.i.d. data)**.** *Let $\widetilde{Y}_1, \ldots, \widetilde{Y}_n$ be independent, mean-zero, sub-exponential random variables and set*

$$S_n := \sum_{i=1}^{n} \widetilde{Y}_i.$$

Then for every $t \geq 0$,

$$\Pr\left(|n^{-1}S_n| \geq t\right) \;\leq\; 2\exp\left(-cn\,\min\left\{\frac{t^2}{\left(\max_i \|\widetilde{Y}_i\|_{\psi_1}\right)^2},\; \frac{t}{\max_i \|\widetilde{Y}_i\|_{\psi_1}}\right\}\right), \quad (35)$$

where $c > 0$ is an absolute constant and

$$\|X\|_{\psi_1} \;:=\; \inf\left\{s > 0 : \mathbb{E}\exp\left(|X|/s\right) \leq 2\right\}$$

denotes the sub-exponential (Orlicz) norm of a real random variable X.

Lemma 2 (Dvoretzky-Kiefer-Wolfowitz inequality)**.** *Let $Y_1, \ldots, Y_n$ be real-valued independent random variables with cumulative distribution function $F(.)$. Further denote $F_n(.)$ the empirical distribution function defined by*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}, \quad x \in \mathbb{R} \quad (36)$$

then for every $t > 0$,

$$\Pr\left(\sup_{y\in\mathbb{R}} |F_n(y) - F(y)| > t\right) \leq 2\exp\left(-2nt^2\right). \quad (37)$$

Lemma 3 (Hoeffding inequality for Bernoulli random variables)**.** *Let $Z_1, \cdots, Z_n$ be independent Bernoulli(p) random variables with $\hat{p} = \frac{1}{n}\sum Z_i$, Since $0 \leq Z_i \leq 1$, Hoeffding (1963, Theorem 2) for any $t > 0$, gives*

$$\Pr\left(|\hat{p} - p| \geq t\right) \leq \exp\left(-2nt^2\right). \quad (38)$$

A.2 Weakly Dependent Data

In this section, we present the definitions and lemmas relevant to weakly dependent data. The definition of the α-mixing process, as well as the concentration inequalities used to derive nonparametric bounds for weakly dependent data drawn from sub-exponential distributions, are drawn from White (2014), Merlevède et al. (2009), Rio (2000) and Dedecker and Merlevède (2007).

Definition 1 (α-mixing process)**.** *Let the sequence of random variables $\widetilde{Y}_1, \cdots, \widetilde{Y}_n$ be defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, where $\mathcal{F}_t = \sigma(\widetilde{Y}_1, \cdots, \widetilde{Y}_t)$ is the σ-field spanned by $\{\widetilde{Y}_i\}_{i=1}^{t}$. Additionally, let $\mathcal{G}$ and $\mathcal{H}$ be two σ-fields such that $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ and define*

$$\alpha(\mathcal{G}, \mathcal{H}) = \sup_{G\in\mathcal{G}, H\in\mathcal{H}} \{|\Pr(G \cap H) - \Pr(G)\Pr(H)|\} \quad (39)$$

and define the Borel σ-field $\mathcal{B}_1^m = \sigma(\widetilde{Y}_1, \cdots, \widetilde{Y}_m)$ and the α-mixing coefficient $\beta(k)$ as

$$\alpha(k) \equiv \sup_m \alpha(\mathcal{B}_1^m, \mathcal{B}_{m+k}^n) \tag{40}$$

If for the sequence $\{\widetilde{Y}_t\}$, $\alpha(k) \to 0$ as $k \to \infty$, $\widetilde{Y}_t$ is called α-mixing.

Lemma 4 (Bernstein inequality for weakly dependent data)**.** *Let $\widetilde{Y}_1, \ldots, \widetilde{Y}_n$ be mean-zero, real-valued random variables drawn from a subexponential distributions that satisfy the α-mixing condition with exponential decay. Moreover, for any positive M, let $\varphi_M(x) = (x \vee M) \wedge (-M)$ and define V as,*

$$V = \sup_{M \geq 1} \sup_{i > 0} \left(Var(\varphi_M(\widetilde{Y}_i)) + 2 \sum_{j>1} |cov(\varphi_M(\widetilde{Y}_i), \varphi_M(\widetilde{Y}_j))| \right) < \infty. \tag{41}$$

Further, define:

$$S_n := \sum_{i=1}^{n} \widetilde{Y}_i.$$

Then for every $n \geq 4$ and $t > 0$, and for positive constants C_1, C_2, C_3, C_4 depending only on c, γ and γ_1, we have

$$\begin{aligned} \Pr\left(|n^{-1} S_j| \geq t\right) &\leq \Pr\left(\sup_{j \leq n} |n^{-1} S_j| \geq t\right) \\ &\leq n \exp\left(-\frac{(nt)^\gamma}{C_1}\right) + \exp\left(-\frac{(nt)^2}{C_2(1+nV)}\right) \\ &\quad + \exp\left(-\frac{(nt)^2}{C_3 n} \exp\left(\frac{(nt)^{\gamma(1-\gamma)}}{C_4 (\log nt)^\gamma}\right)\right) \end{aligned}$$

Lemma 5 (Dvoretzky-Kiefer-Wolfowitz inequality for weakly dependent data)**.** *Let $Y_1, \cdots, Y_n$ be a strictly stationray real-valued sequence with common CDF. F and assume the strong mixing coefficients $\alpha(k)$ in* (40) *satisfies $\sum_{k \geq 1} \alpha(k)^{1/2} < \infty$. Define the empirical CDF as per Eq.* (36)*. Then for every $t > 0$ and $n \geq 1$*

$$P\left(\sup_{y \in \mathbb{R}} |F_n(y) - F(y)| > t\right) \leq 2 \exp\left(-\frac{nt^2}{2(1+4C_\alpha)^2}\right) \tag{42}$$

where $C_\alpha = \sum_{k \geq 1} \alpha(k)^{1/2} < \infty$. In particular, if $\alpha(k) = 0$ for all $k \geq 1$ (the independent case) then $C_\alpha = 0$ and (47) *reduces to* (37)*.*

Lemma 6 (Hoeffding inequality for α-mixing Bernoulli data)**.** *Let $Z_1, \cdots, Z_n$ be a strictly stationary $\{0,1\}$-valued sequence with $p = \mathbb{E}[Z_1]$ and $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n Z_i$, and strong-mixing coefficients $\alpha(k)$. Assume $C_\alpha = \sum_{k \geq 1} \alpha(k)^{1/2} < \infty$. Then for every $t > 0$ and $n \geq 1$,*

$$\Pr\left(|\hat{p}_n - p| \geq t\right) \leq 2 \exp\left(-\frac{n\, t^2}{2(1+4C_\alpha)^2}\right). \tag{43}$$

In particular, if $\alpha(k) \equiv 0$ then $C_\alpha = 0$ and this reduces to the usual Azuma–Hoeffding bound $\Pr(|\hat{p}_n - p| \geq t) \leq 2e^{-nt^2/2}$.

B Proofs

B.1 Proof of Lemma 5

From Section 2, Theorem 1 and Remark 1 of Dedecker and Merlevède (2007), it is known that for any finite measure μ and $p \geq 2$:

$$\Pr\left(\sqrt{n}\|F_n - F\|_{p,\mu} \geq x\right) \leq 2\exp\left(-\frac{x^2}{2(p-1)(\|Z_1\|_{p,\mu} + 2\sum_{k\geq 1}\tau_{\mu,p,\infty}(k))^2}\right) \tag{44}$$

where $Z_i(t) = \mathbb{1}\{X_i \leq t\} - F(t)$ and $\tau_{p,\mu,\infty} = \left\|\|\mathbb{E}\left(Z_{k+1} \mid \mathcal{M}_0\right)\|_{p,\mu}\right\|_\infty$. By choosing the Kolmogorov norm, i.e., setting $p = 2$ and $\mu = \lambda_1$ (Lebesgue measure on $[0,1]$) in Eq. (44), we obtain the deviation bound for Kolmogorov distance $\sup_x|F_n - F|$.

Next we relate the τ coefficients to α-mixing. Inequality (4.1) in Section 4.1 of Dedecker and Merlevède (2007) shows:

$$\tau_{\lambda_1,2,1}(k) \leq 18\alpha(k). \tag{45}$$

Since $\tau_{\lambda_1,2,\infty}(k) \leq \tau_{\lambda_1,2,1}(k)^{1/2}$, we get

$$\tau_{\lambda_1,2,\infty}(k) \leq 18^{1/2}\alpha(k)^{1/2}. \tag{46}$$

After minor algebra, taking $x = \sqrt{n}t$ and recalling $\|Z_1\|_{2,\lambda_1} \leq 1$, we arrive at

$$P\left(\sup_{y\in\mathbb{R}}|F_n(y) - F(y)| > t\right) \leq 2\exp\left(-\frac{nt^2}{2(1+4C_\alpha)^2}\right) \tag{47}$$

with

$$1 + 4\sum_{k\geq 1}\tau_{\lambda_1,2,\infty}(k) \leq 1 + 4(18)^{1/2}\sum_{k\geq 1}\alpha(k)^{1/2} = 1 + 4C_\alpha.$$

B.2 Proof of Proposition 1

The theory that we have laid out thus far concerns the identification problem. However, empirical research must also be concerned with sampling variation. Note that the empirical counterpart of the nonparametric bound (14) is:

$$\Re \in \left[\hat{\delta}_1\hat{p}_1 + L^{(1)}\hat{p}_0 - U^{(0)}\hat{p}_1 - \hat{\delta}_0\hat{p}_0,\ \hat{\delta}_1\hat{p}_1 + U^{(1)}\hat{p}_0 - L^{(0)}\hat{p}_1 - \hat{\delta}_0\hat{p}_0\right] \tag{48}$$

For $u = m_0, \cdots, m_1$, to simultaneously obtain the $(1-\alpha_u)\%$ confidence set for both the upper and lower bounds for the identification region (48), we must first find the confidence bounds with an appropriate significance level and combine them using Bonferroni inequalities, such that the combined confidence set has a $(1-\alpha_u)\%$ coverage rate, or:

$$\Pr\Big([\mathcal{L}(\hat{\theta}), \mathcal{U}(\hat{\theta})] \subseteq \mathbb{I}(\Re)\Big) \geq 1 - \alpha_u, \quad \text{with} \quad \alpha_u = \frac{\alpha}{\overline{\mathcal{M}}}. \tag{49}$$

where $\hat{\theta} = (\hat{\delta}_1, \hat{\delta}_0, \hat{p}_1, \hat{p}_0)^\top$ is a 4×1 vector of estimators and $\mathbb{I}(\Re)$ is an interval $[l(\Re), u(\Re)]$. In other words, we wish to obtain:

$$\Pr\Big(l(\Re) \leq \mathcal{L}(\hat{\theta})\Big) \geq 1 - \frac{\alpha_u}{2}, \quad \text{and} \quad \Pr\Big(u(\Re) \geq \mathcal{U}(\hat{\theta})\Big) \geq 1 - \frac{\alpha_u}{2} \tag{50}$$

such that,

$$\Pr\left(l(\Re) \leq \mathcal{L}(\hat{\theta}) \cap u(\Re) \geq \mathcal{U}(\hat{\theta})\right) \geq 1 - \alpha_u \tag{51}$$

We know from Boole's inequality that:

$$\begin{aligned}\Pr\left(l(\Re) \leq \mathcal{L}(\hat{\theta}) \cap u(\Re) \geq \mathcal{U}(\hat{\theta})\right) &\geq 1 - \Pr\left(l(\Re) > \mathcal{L}(\hat{\theta})\right) - \Pr\left(u(\Re) < \mathcal{U}(\hat{\theta})\right) \\ &\geq 1 - \frac{\alpha_u}{2} - \frac{\alpha_u}{2} \\ &\geq 1 - \alpha_u\end{aligned} \tag{52}$$

where the significance levels are such that $1 - \alpha_{u,1} - \alpha_{u,2} = 1 - \alpha_u/2 - \alpha_u/2 = 1 - 0.05$. Thus,

$$\begin{aligned}\mathcal{L}(\hat{\theta}) - \Phi^{-1}(1-\alpha_u/2) \times \text{S.E.}\left(\mathcal{L}(\hat{\theta})\right) \leq \mathcal{L}(\theta) \leq \mathcal{L}(\hat{\theta}) + \Phi^{-1}(1-\alpha_u/2) \times \text{S.E.}\left(\mathcal{L}(\hat{\theta})\right) \\ \mathcal{U}(\hat{\theta}) - \Phi^{-1}(1-\alpha_u/2) \times \text{S.E.}\left(\mathcal{U}(\hat{\theta})\right) \leq \mathcal{U}(\theta) \leq \mathcal{U}(\hat{\theta}) + \Phi^{-1}(1-\alpha_u/2) \times \text{S.E.}\left(\mathcal{U}(\hat{\theta})\right)\end{aligned} \tag{53}$$

It remains to find the standard errors of $\mathcal{L}(\hat{\theta})$ and $\mathcal{U}(\hat{\theta})$, which is a rather tedious task due to the nonlinear nature of the estimators. Given the relatively large sample sizes, we may rely on the delta method.

By definition, the consistent estimator $\hat{\theta}$ converges in probability to its true value θ, and the CLT can be applied to obtain asymptotic normality, i.e.,

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{D} N(0, \Omega_\theta) \tag{54}$$

By Taylor expansion of $\mathcal{L}(\hat{\theta})$ and $\mathcal{U}(\hat{\theta})$:

$$\mathcal{L}(\hat{\theta}) \approx \mathcal{L}(\theta) + \nabla\mathcal{L}(\theta)^\top(\hat{\theta} - \theta) \tag{55}$$

$$\mathcal{U}(\hat{\theta}) \approx \mathcal{U}(\theta) + \nabla\mathcal{U}(\theta)^\top(\hat{\theta} - \theta) \tag{56}$$

where

$$\nabla\mathcal{L}(\theta) = \left(\frac{\partial}{\partial\delta_1}\mathcal{L}(\theta), \frac{\partial}{\partial\delta_0}\mathcal{L}(\theta), \frac{\partial}{\partial p_1}\mathcal{L}(\theta), \frac{\partial}{\partial p_0}\mathcal{L}(\theta)\right)^\top \tag{57}$$

with $\nabla\mathcal{U}(\theta)$ defined similarly. Therefore, taking the Var(.) of both sides of the equations (55) and (56) yields:

$$\begin{aligned}\text{Var}\left(\mathcal{L}(\hat{\theta})\right) &\approx \text{Var}\left(\mathcal{L}(\theta) + \nabla\mathcal{L}(\theta)^\top(\hat{\theta} - \theta)\right) \\ &= \text{Var}\left(\mathcal{L}(\theta) + \nabla\mathcal{L}(\theta)^\top\hat{\theta} - \nabla\mathcal{L}(\theta)^\top\theta)\right) \\ &= \text{Var}\left(\nabla\mathcal{L}(\theta)^\top\hat{\theta}\right) \\ &= \nabla\mathcal{L}(\theta)^\top cov\left(\hat{\theta}\right)\nabla\mathcal{L}(\theta) \\ &= \nabla\mathcal{L}(\theta)^\top\frac{\Omega_0}{N}\nabla\mathcal{L}(\theta)\end{aligned} \tag{58}$$

with $\text{Var}\left(\mathcal{L}(\hat{\theta})\right)$ defined similarly. We know, that

$$\nabla\mathcal{L}(\theta) = \left(p_1, -p_0, \delta_1 - U^{(0)}, L^{(1)} - \delta_0\right)^\top \tag{59}$$

$$\nabla\mathcal{U}(\theta) = \left(p_1, -p_0, \delta_1 - L^{(0)}, U^{(1)} - \delta_0\right)^\top \tag{60}$$

The covariance matrix of estimators $\hat{\theta}$ is given explicitly by:

$$\Omega_{\hat{\theta}} = \begin{pmatrix} \text{Var}(\hat{\delta}_1) & 0 & 0 & 0 \\ 0 & \text{Var}(\hat{\delta}_0) & 0 & 0 \\ 0 & 0 & \text{Var}(\hat{p}_1) & -\text{Var}(\hat{p}_1) \\ 0 & 0 & -\text{Var}(\hat{p}_1) & \text{Var}(\hat{p}_0) \end{pmatrix}, \tag{61}$$

Therefore,

$$\sqrt{N}\left(\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta)\right) \xrightarrow{D} N\left(0, \nabla\mathcal{L}(\theta)^\top \Omega \nabla\mathcal{L}(\theta)\right) \tag{62}$$

$$\sqrt{N}\left(\mathcal{U}(\hat{\theta}) - \mathcal{U}(\theta)\right) \xrightarrow{D} N\left(0, \nabla\mathcal{U}(\theta)^\top \Omega \nabla\mathcal{U}(\theta)\right) \tag{63}$$

B.3 Proof of Proposition 2

We wish to show how to obtain the coverage probability

$$\Pr\left(\forall u \in \mathcal{M}, \Re_u \in H_{\alpha_u}[\Re_u]\right) \geq 1 - \alpha \tag{64}$$

for $u = m_0, \cdots, m_1$ and some arbitrary significance level $0 < \alpha < 1$ when assumption 1 is in place. To achieve this, we first need to consider the six "good" events:

$$\begin{aligned}
\mathcal{E}_1 &:= \{|\hat{\mu}_1 - \mu_1| \leq t_1\} \\
\mathcal{E}_2 &:= \{|\hat{\mu}_0 - \mu_0| \leq t_2\} \\
\mathcal{E}_3 &:= \{|\hat{p}_1 - p_1| \leq t_3\} \\
\mathcal{E}_4 &:= \{|\hat{p}_0 - p_0| \leq t_4\} \\
\mathcal{E}_5 &:= \left\{\sup_y \left|F_{N_1}^{(1)}(y) - F^{(1)}(y)\right| \leq t_5\right\} \\
\mathcal{E}_6 &:= \left\{\sup_y \left|F_{N_0}^{(0)}(y) - F^{(0)}(y)\right| \leq t_6\right\}
\end{aligned}$$

Thus, showing $\Pr\left(\Re_u \in H_{\alpha_u}[\Re_u]\right) \geq 1 - \alpha_u$ for $u = m_0, \cdots, m_1$ is equivalent to showing that the intersection of the events, i.e., $\Pr(\bigcap_{i=1}^{6} \mathcal{E}_i) \geq 1 - \alpha_u$. Using De Morgan's law, it is clear that

$$\Pr\left(\bigcap_{i=1}^{6} \mathcal{E}_i\right) = 1 - \Pr\left(\bigcup_{i=1}^{6} \mathcal{E}_i^c\right) \tag{65}$$

where $\mathcal{E}_i^c$ is the complement of the event $\mathcal{E}_i$. Furthermore, we know from Boole's inequality that

$$\Pr\left(\bigcup_{i=1}^{6} \mathcal{E}_i^c\right) \leq \sum_{i=1}^{6} \Pr\left(\mathcal{E}_i^c\right). \tag{66}$$

Consequently,

$$\Pr\left(\bigcap_{i=1}^{6}\mathcal{E}_i\right) = 1 - \Pr\left(\bigcup_{i=1}^{6}\mathcal{E}_i^c\right) \tag{67}$$

$$\geq 1 - \sum_{i=1}^{6}\Pr\left(\mathcal{E}_i^c\right) \tag{68}$$

Hence, showing that the bound (64) holds is equivalent to ensuring that $\sum_{i=1}^{6}\Pr\left(\mathcal{E}_i^c\right) \leq \alpha_u$ for $u = m_0, \cdots, m_1$. The only tools we need are the three inequalities in Lemmas 1–3.

(i) Means μ_1, μ_0 (events $\mathcal{E}_1, \mathcal{E}_2$). Let N_1 (resp. N_0) be the number of observations with $Z = 1$ (resp. $Z = 0$). Lemma 1 gives for any $t > 0$

$$\Pr\left(|\hat{\mu}_k - \mu_k| \geq t\right) \leq 2\exp\left[-cN_k \min\left\{t^2/M_k^2, t/M_k\right\}\right], \qquad k = 0, 1,$$

where $M_k := \max_{i:Z_i=k}\|Y_i^{(k)} - \mu_k\|_{\psi_1}$. Choose for each arm

$$t_k := \min\left\{M_k\sqrt{\frac{\log(12/\alpha_u)}{cN_k}}, \frac{M_k}{cN_k}\log\left(\frac{12}{\alpha_u}\right)\right\}, \qquad k = 0, 1. \tag{69}$$

(The first term is used when $t_k \leq M_k$ — the "quadratic" regime; otherwise the second, "linear", term is smaller.) With this choice $2\exp[-\log(12/\alpha_u)] = \alpha_u/6$, so $\Pr(\mathcal{E}_1^c) = \Pr(\mathcal{E}_2^c) = \alpha_u/6$.

(ii) Treatment proportions p_1, p_0 (events $\mathcal{E}_3, \mathcal{E}_4$). With $N = N_1 + N_0$, Lemma 3 yields

$$\Pr\left(|\hat{p}_k - p_k| \geq t\right) \leq 2\exp[-2Nt^2], \qquad k = 0, 1.$$

Set

$$t_3 = t_4 := \sqrt{\frac{\log(12/\alpha_u)}{2N}}, \tag{70}$$

so that $\Pr(\mathcal{E}_3^c) = \Pr(\mathcal{E}_4^c) = \alpha_u/6$.

(iii) Empirical CDFs (events $\mathcal{E}_5, \mathcal{E}_6$). Lemma 2 (two-sided DKW) gives

$$\Pr\left(\sup_y |F_n^{(k)}(y) - F^{(k)}(y)| > t\right) \leq 2\exp[-2N_k t^2], \qquad k = 0, 1.$$

Choose

$$t_5 := \sqrt{\frac{\log(12/\alpha_u)}{2N_1}}, \qquad t_6 := \sqrt{\frac{\log(12/\alpha_u)}{2N_0}}, \tag{71}$$

so that $\Pr(\mathcal{E}_5^c) = \Pr(\mathcal{E}_6^c) = \alpha_u/6$.

Step 1 concluded. By construction, $\Pr(\mathcal{E}_i^c) \le \alpha_u/6$ for each i, hence

$$\Pr\left(\bigcap_{i=1}^{6} \mathcal{E}_i\right) \ge 1 - \sum_{i=1}^{6} \Pr(\mathcal{E}_i^c) \ge 1 - \alpha_u. \tag{72}$$

Step 2. From the intersection event to coverage. On $\cap_{i=1}^{6}\mathcal{E}_i$ we have $|\hat{\mu}_k - \mu_k| \le t_k$, $|\hat{p}_k - p_k| \le t_{k+2}$, and $\sup_y |F_n^{(k)}(y) - F^{(k)}(y)| \le t_{k+4}$ for $k = 0, 1$. These inequalities imply $L^{(k)}(t_{k+4}) \le \mathbb{E}[Y^{(k)} \mid Z = k] \le U^{(k)}(t_{k+4})$ with $L^{(k)}(t) := Y_{(1)}^{(k)} - t$ and $U^{(k)}(t) := Y_{(n_k)}^{(k)} + t$. Plugging the modified support bounds, the perturbed means, and the perturbed treatment shares into Manski's lower and upper expressions yields two numbers $L_\alpha(\hat{\theta}) \le U_\alpha(\hat{\theta})$ such that $\Re \in [L_{\alpha_u}(\hat{\theta}), U_{\alpha_u}(\hat{\theta})]$ whenever $(\hat{\theta}, Y) \in \cap_{i=1}^{6}\mathcal{E}_i$. Consequently,

$$\Pr\left(\Re_u \in H_{\alpha_u}[\Re_u]\right) \ge \Pr\left(\bigcap_{i=1}^{6} \mathcal{E}_i\right) \overset{(72)}{\ge} 1 - \alpha_u, \quad \text{for} \quad u = m_0, \cdots, m_1$$

This establishes (64).

B.4 Proof of Corollary 1

The argument follows Proposition 2 verbatim except that the empirical–CDF events are now one-sided because the lower support is the known constant λ:

$$\mathcal{E}_5 := \left\{ \sup_y \left(F_{N_1}^{(1)}(y) - F^{(1)}(y) \right) \le t_5 \right\},$$
$$\mathcal{E}_6 := \left\{ \sup_y \left(F_{N_0}^{(0)}(y) - F^{(0)}(y) \right) \le t_6 \right\}.$$

For a one-sided Kolmogorov deviation the DKW inequality is

$$\Pr\left(\sup_y \left[F_n(y) - F(y) \right] > t \right) \le \exp\left(-2nt^2\right),$$

so choosing

$$t_5 := \sqrt{\frac{\log(6/\alpha_u)}{2N_1}}, \qquad t_6 := \sqrt{\frac{\log(6/\alpha_u)}{2N_0}}$$

ensures $\Pr(\mathcal{E}_5^c) = \Pr(\mathcal{E}_6^c) \le \alpha_u/6$. The four mean- and share-events $\mathcal{E}_1$–$\mathcal{E}_4$ and their bounds are unchanged, hence each still receives probability $\alpha_u/6$. Because the six complements jointly spend at most α_u, Boole's inequality and the algebra in Proposition 2 give

$$\Pr\left(\Re_u \in H_{\alpha_u}[\Re_u]\right) \ge 1 - \alpha_u, \qquad u = m_0, \ldots, m_1.$$

B.5 Proof of Proposition 3

Similar to the proof of Proposition 2, define the six "good" events $\mathcal{E}_1 \cdots, \mathcal{E}_6$ and apply Boole's (union) bound. We then choose each threshold t_i so that $\Pr(\mathcal{E}_i^c) \le \alpha_u/6$ for $u = m_0, \cdots, m_1$ under dependence:

(i) Means μ_1, μ_0 (events $\mathcal{E}_1, \mathcal{E}_2$). By Lemma 4,

$$\Pr\left(|\hat{\mu}_k - \mu_k| \geq t_k\right) \leq T_1(t_k) + T_2(t_k) + T_3(t_k),$$

where

$$T_1(t) = N_k \exp\left(-\frac{(N_k\, t)^\gamma}{C_1}\right), \qquad \text{(term 1)}$$

$$T_2(t) = \exp\left(-\frac{(N_k\, t)^2}{C_2(1 + N_k V)}\right), \qquad \text{(term 2)}$$

$$T_3(t) = \exp\left(-\frac{(N_k\, t)^2}{C_3\, N_k} \exp\left(\frac{(N_k\, t)^{\gamma(1-\gamma)}}{C_4(\log(N_k t))^\gamma}\right)\right). \qquad \text{(term 3)}$$

To make each term $\leq \alpha_u/18$ for $u = m_0, \cdots, m_1$:

1. $T_1(t) \leq \alpha_u/18$ iff

$$\frac{(N_k t)^\gamma}{C_1} \geq \log\frac{18 N_k}{\alpha_u} \quad \Longrightarrow \quad t \geq t_k^{(1)} := \frac{(C_1 \log(18 N_k/\alpha_u))^{1/\gamma}}{N_k}.$$

2. $T_2(t) \leq \alpha_u/18$ iff

$$\frac{(N_k t)^2}{C_2(1 + N_k V)} \geq \log\frac{18}{\alpha_u} \quad \Longrightarrow \quad t \geq t_k^{(2)} := \frac{\sqrt{C_2(1 + N_k V)\log(18/\alpha_u)}}{N_k}.$$

3. $T_3(t) \leq \alpha_u/18$ iff

$$\frac{(N_k t)^2}{C_3 N_k} \exp\left(\frac{(N_k t)^{\gamma(1-\gamma)}}{C_4(\log(N_k t))^\gamma}\right) \geq \log\frac{18}{\alpha_u},$$

so define $t_k^{(3)}$ to be the unique positive solution of this equation. Finally set

$$t_k = \max\{t_k^{(1)}, t_k^{(2)}, t_k^{(3)}\},$$

then $\Pr(\mathcal{E}_k^c) \leq 3 \cdot (\alpha_u/18) = \alpha_u/6$.

(ii) Treatment proportions p_1, p_0 (events $\mathcal{E}_3, \mathcal{E}_4$). By Lemma 6,

$$\Pr\left(|\hat{p}_k - p_k| \geq t_k\right) \leq 2\exp\left(-\frac{N_k\, t_k^2}{2(1 + 4C_\alpha)^2}\right).$$

Solving $2\exp(-A) \leq \alpha_u/6$ with $A = N_k\, t^2/[2(1 + 4C_\alpha)^2]$ gives

$$t_3 = t_4 = (1 + 4C_\alpha)\sqrt{\frac{2\log(12/\alpha_u)}{N_k}},$$

so $\Pr(\mathcal{E}_3^c) = \Pr(\mathcal{E}_4^c) = \alpha_u/6$.

(iii) Empirical CDFs (events $\mathcal{E}_5, \mathcal{E}_6$). By Lemma 5,

$$\Pr\left(\sup_y \left|F_{N_k}^{(k)}(y) - F^{(k)}(y)\right| > t_k\right) \leq 2\exp\left(-\frac{N_k, t_k^2}{2(1+4C_\alpha)^2}\right).$$

Similarly, set

$$t_5 = t_6 = (1+4C_\alpha)\sqrt{\frac{2\log(12/\alpha_u)}{N_k}},$$

so $\Pr(\mathcal{E}_5^c) = \Pr(\mathcal{E}_6^c) = \alpha_u/6$.

Combining these gives

$$\Pr\left(\bigcap_{i=1}^{6}\mathcal{E}_i\right) \geq 1 - \sum_{i=1}^{6}\Pr(\mathcal{E}_i^c) = 1 - \alpha_u.$$

On $\bigcap_i \mathcal{E}_i$, elementary algebra shows $\Re_u \in H_{\alpha_u}[\Re_u]$, so

$$\Pr\left(\forall u \in \mathcal{M}, \Re_u \in H_{\alpha_u}[\Re_u]\right) \geq 1 - \alpha.$$

This completes the proof.

B.6 Proof of Proposition 4

Let $(Y_{ijt}, Z_{ijt})_{t=1}^T$ be strictly stationary and $\boldsymbol{\alpha}$–mixing with $C_\alpha = \sum_{r\geq 1}\boldsymbol{\alpha}(r)^{1/2} < \infty$. To simultaneously obtain a the $100(1-\alpha)\%$ confidence set for both the upper and lower bounds for the identification region (48), we must first find the confidence bounds with an appropriate significance level and combine them using Bonferroni inequalities, such that the combined confidence set has a $(1-\alpha)\%$ coverage rate, or:

$$\Pr\Bigg(\forall u \in \mathcal{M}, \left\{[\mathcal{L}(\hat{\theta}), \mathcal{U}(\hat{\theta})] \subseteq \mathbb{I}(\Re_u)\right\} \cap \left\{\sup_y \left|F_{N_0}^{(0)}(y) - F^{(0)}(y)\right| \leq \epsilon_0\right\} \\ \cap \left\{\sup_y \left|F_{N_1}^{(1)}(y) - F^{(1)}(y)\right| \leq \epsilon_1\right\}\Bigg) \geq 1 - \alpha, \tag{73}$$

where $\hat{\theta} = (\hat{\delta}_1, \hat{\delta}_0, \hat{p}_1, \hat{p}_0)^\top$ is a 4×1 vector of estimators and $\mathbb{I}(\Re_u)$ is an interval $[l(\Re_u), u(\Re_u)]$. In other words, for $u = m_0, \cdots, m_1$, we wish to obtain:

$$\Pr\left(l(\Re_u) \leq \mathcal{L}(\hat{\theta})\right) \geq 1 - \frac{\alpha_u}{4}, \qquad \Pr\left(u(\Re_u) \geq \mathcal{U}(\hat{\theta})\right) \geq 1 - \frac{\alpha_u}{4}$$
$$\Pr\left(\sup_y \left|F_{N_k}^{(k)}(y) - F^{(k)}(y)\right| \leq \epsilon^{(k)}\right) \geq 1 - \frac{\alpha_u}{4}, \quad \text{for} \quad k = 0, 1.$$

Let us define the events:

$$\mathcal{E}_1 = \{l(\Re_u) \geq \mathcal{L}(\hat{\theta})\}$$
$$\mathcal{E}_2 = \{u(\Re_u) \leq \mathcal{U}(\hat{\theta})\}$$
$$\mathcal{E}_3 = \left\{\sup_y \left|F_{N_1}^{(1)}(y) - F^{(1)}(y)\right| \leq \epsilon_1\right\}$$
$$\mathcal{E}_4 = \left\{\sup_y \left|F_{N_0}^{(0)}(y) - F^{(0)}(y)\right| \leq \epsilon_0\right\}$$

We know from Boole's inequality that:

$$\begin{aligned}\Pr\left(\bigcap_{i=1}^{4}\mathcal{E}_i\right) &\geq 1-\sum_{i=1}^{4}\Pr\left(\mathcal{E}_i^c\right)\\ &\geq 1-4\frac{\alpha_u}{4}\\ &\geq 1-\alpha_u.\end{aligned} \tag{74}$$

Thus,

$$\begin{aligned}\mathcal{L}(\hat{\theta})-\Phi^{-1}(1-\alpha_u/4)\times\text{S.E.}\left(\mathcal{L}(\hat{\theta})\right)\leq\mathcal{L}(\theta)\leq\mathcal{L}(\hat{\theta})+\Phi^{-1}(1-\alpha_u/4)\times\text{S.E.}\left(\mathcal{L}(\hat{\theta})\right)\\ \mathcal{U}(\hat{\theta})-\Phi^{-1}(1-\alpha_u/4)\times\text{S.E.}\left(\mathcal{U}(\hat{\theta})\right)\leq\mathcal{U}(\theta)\leq\mathcal{U}(\hat{\theta})+\Phi^{-1}(1-\alpha_u/4)\times\text{S.E.}\left(\mathcal{U}(\hat{\theta})\right)\end{aligned} \tag{75}$$

Additionally, we must ensure that:

$$\begin{aligned}\Pr\left(\sup_y\left|F_{N_k}^{(k)}(y)-F^{(k)}(y)\right|>\epsilon_k\right)&\geq 2\exp\left(-\frac{N_k\epsilon_k^2}{2(1+4C_\alpha)^2}\right)\\ &=\frac{\alpha_u}{4}\quad\text{for}\quad k=0,1\end{aligned}$$

which holds when

$$\epsilon_k:=(1+4C_\alpha)\sqrt{\frac{2\log(8/\alpha_u)}{N_k}}. \tag{76}$$

It remains to find the standard errors of $\mathcal{L}(\hat{\theta})$ and $\mathcal{U}(\hat{\theta})$, which is a rather tedious task due to the nonlinear nature of the estimators. Given the relatively large sample sizes, we may rely on the delta method.

By definition, the consistent estimator $\hat{\theta}$ converges in probability to its true value θ, and the CLT can be applied to obtain asymptotic normality, i.e.,

$$\sqrt{N}(\hat{\theta}-\theta)\xrightarrow{D}N(0,\Omega_\theta). \tag{77}$$

From the proof of Proposition 1 and given the sample size in treatment group $k\in\{0,1\}$ and by

$$Y_{(1)}^{(k)}\leq\cdots\leq Y_{(N_k)}^{(k)},$$

we know that

$$\nabla\mathcal{L}(\theta)=\left(p_1,-p_0,\delta_1-U_{\alpha_u}^{(0)},L_{\alpha_u}^{(1)}-\delta_0\right)^\top \tag{78}$$

$$\nabla\mathcal{U}(\theta)=\left(p_1,-p_0,\delta_1-L_{\alpha_u}^{(0)},U_{\alpha_u}^{(1)}-\delta_0\right)^\top \tag{79}$$

where unlike in Proposition 1,

$$L_{\alpha_u}^{(k)}:=Y_{(1)}^{(k)}-\epsilon_k\quad\text{and}\quad U_{\alpha_u}^{(k)}:=Y_{(N_k)}^{(k)}+\epsilon_k \tag{80}$$

where ϵ_k is defined in Eq. (76). Similarly, the covariance matrix of estimators $\hat{\theta}$ is given explicitly by:

$$\Omega_{\hat{\theta}}=\begin{pmatrix}\text{Var}(\hat{\delta}_1)&0&0&0\\0&\text{Var}(\hat{\delta}_0)&0&0\\0&0&\text{Var}(\hat{p}_1)&-\text{Var}(\hat{p}_1)\\0&0&-\text{Var}(\hat{p}_1)&\text{Var}(\hat{p}_0)\end{pmatrix}, \tag{81}$$

Therefore,

$$\sqrt{N}\left(\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta)\right) \xrightarrow{D} N\left(0, \nabla\mathcal{L}(\theta)^\top \Omega \nabla\mathcal{L}(\theta)\right) \tag{82}$$

$$\sqrt{N}\left(\mathcal{U}(\hat{\theta}) - \mathcal{U}(\theta)\right) \xrightarrow{D} N\left(0, \nabla\mathcal{U}(\theta)^\top \Omega \nabla\mathcal{U}(\theta)\right) \tag{83}$$

C Algorithms

Algorithm 1 Monte-Carlo experiment: single–threshold Manski vs. Hybrid band

Fixed inputs

Replications $B = 2{,}000$; firms $n = 50$; panel lengths $T \in \{1, 2, 5\}$;
diversity cut-off $\tau^\circ = 50\%$ (no grid); overall size $\alpha = 0.05$;
constant treatment effect $\Delta = 4$; two-sided critical value $z_{0.975} = 1.96$.

Step 0 (once per design)

1. *Latent support.*
   - Design F: known lower bound $a^* = 0$; draw an oracle sample of size N_{big} and set $b^* = \max Y_{it}^0 + \Delta$.
   - Design G: published support $(a^*, b^*) = (-5, 5)$.
   - Designs A–E: oracle sample as above and set $a^* = \min Y_{it}^0$, $b^* = \max Y_{it}^0 + \Delta$.

for $b = 1$ **to** B **do** ▷ replication loop
  Draw a fresh baseline $\{Y_{it}^0, D_{it}\}_{i=1,\dots,n;\, t=1,\dots,T}$.
  $Y_{it} = Y_{it}^0 + \Delta D_{it}$.
  Empirical extrema $\widehat{a} = \min Y_{it}^0$, $\widehat{b} = \max Y_{it}^0$.
  $$(a, b) \leftarrow \begin{cases} (a^*, b^*), & \text{design G (both bounds known)}, \\ (0, \widehat{b}), & \text{design F (lower bound known)}, \\ (\widehat{a}, \widehat{b}), & \text{designs A–E (unknown support)}. \end{cases}$$
  Compute $(L^{\mathrm{M}}, U^{\mathrm{M}})$ using (a, b).
  if design G **then**
    $(L^{\mathrm{H}}, U^{\mathrm{H}}) \leftarrow (L^{\mathrm{M}}, U^{\mathrm{M}})$
  else
    $$\log C \leftarrow \begin{cases} \log(1/\alpha), & \text{design F (one-sided DKW)}, \\ \log(2/\alpha), & \text{designs A–E (two-sided DKW)}. \end{cases}$$
    $\varepsilon_n \leftarrow \sqrt{\log C/(2nT)}$
    $(\hat{\sigma}_L, \hat{\sigma}_U) \leftarrow$ delta-method SEs
    $(L^{\mathrm{H}}, U^{\mathrm{H}}) \leftarrow \left[L^{\mathrm{M}} - \varepsilon_n - z_{0.9875}\hat{\sigma}_L,\ U^{\mathrm{M}} + \varepsilon_n + z_{0.9875}\hat{\sigma}_U\right]$
  end if
  $\mathrm{hit}_{\mathrm{M}}[b] \leftarrow \mathbb{1}\{\Delta \in [L^{\mathrm{M}}, U^{\mathrm{M}}]\}$
  $\mathrm{hit}_{\mathrm{H}}[b] \leftarrow \mathbb{1}\{\Delta \in [L^{\mathrm{H}}, U^{\mathrm{H}}]\}$
end for
Output: empirical coverages
$\widehat{P}_{\mathrm{M}} = B^{-1} \sum_b \mathrm{hit}_{\mathrm{M}}[b]$, $\widehat{P}_{\mathrm{H}} = B^{-1} \sum_b \mathrm{hit}_{\mathrm{H}}[b]$.

D Additional Analysis

D.1 Sector Group Classifications

Group	Included Sectors
Cyclicals	Consumer Discretionary, Materials, Industrials, Real Estate
Defensives	Health Care, Consumer Staples, Utilities
Growth & Innovation	Information Technology, Communication Services
Financials	Financials
Energy	Energy

Table D.1: Sector Group Classifications

D.2 Clustered Descriptive Statistics

This section presents the descriptive statistics and kernel density plots for the sectoral groups described in Table D.1, and for all companies prior to 1st September 2019.

Table D.2: Descriptive Statistics - (Pre 01/09/2019)

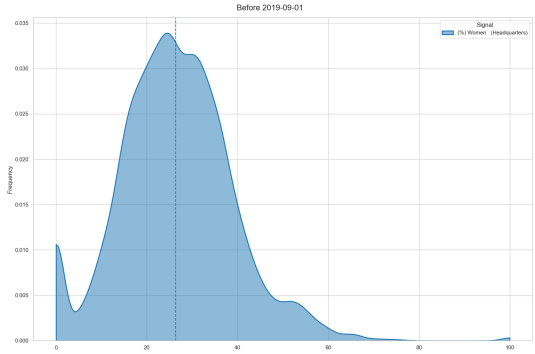
Variable	Min	Mean	Median	Max	Std Dev	Skewness	Kurtosis	N
(%) Women	0.000	26.340	25.902	100	12.710	0.449	1.707	16066
(%) Unknown gender	0.000	0.028	0.022	0.496	0.032	2.713	18.328	16066
Tobin's Q	-0.612	0.386	0.020	5.047	1.094	2.218	5.456	14466
Total assets	10.392	16.014	16.037	21.740	1.840	0.044	0.187	15126
Leverage	0.000	0.288	0.272	3.945	0.229	3.677	41.331	15119
Total employees	85.559	24312.371	7951.079	703268.060	49761.515	5.480	43.774	16214

Table D.3: Descriptive Statistics - (Cyclicals)

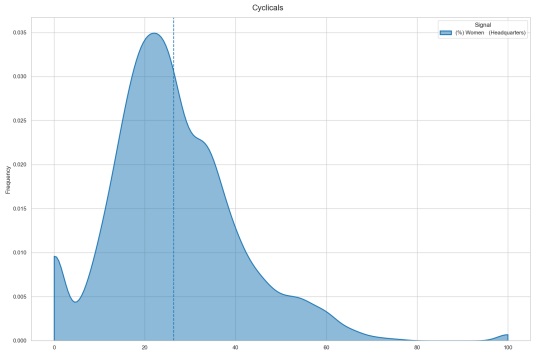
Variable	Min	Mean	Median	Max	Std Dev	Skewness	Kurtosis	N
(%) Women	0.000	26.359	24.544	100	14.424	0.896	2.384	10429
(%) Unknown gender	0.000	0.023	0.015	0.331	0.030	2.933	16.268	10429
Tobin's Q	-0.612	0.321	0.045	5.047	0.880	2.494	8.140	9833
Total assets	11.064	15.619	15.798	20.230	1.453	-0.298	-0.020	10120
Leverage	0.000	0.347	0.327	3.945	0.257	4.691	47.105	10118
Total employees	148.263	20767.166	8964.840	941046.440	40760.375	8.815	137.453	10556

Table D.4: Descriptive Statistics - (Defensives)

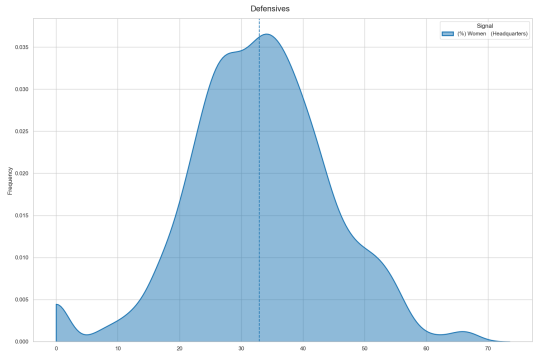
Variable	Min	Mean	Median	Max	Std Dev	Skewness	Kurtosis	N
(%) Women	0.000	32.928	33.008	67.221	11.497	-0.183	0.653	4953
(%) Unknown gender	0.000	0.041	0.036	0.496	0.037	2.398	17.184	4953
Tobin's Q	-0.612	0.582	0.103	5.047	1.263	1.908	3.286	4712
Total assets	10.632	16.308	16.478	19.347	1.545	-0.500	-0.185	4860
Leverage	0.000	0.344	0.337	2.013	0.177	0.903	5.015	4858
Total employees	99.419	24269.252	7813.385	430494.690	44319.281	4.504	28.430	4956



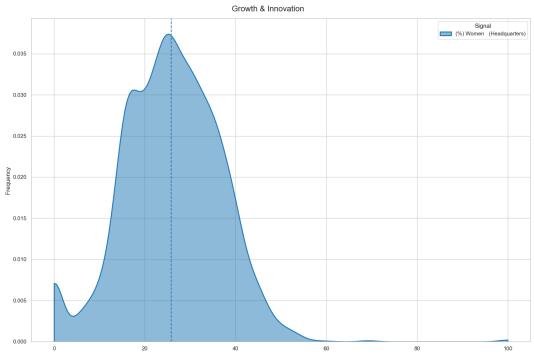
(a) All firms (before 01-09-2019)



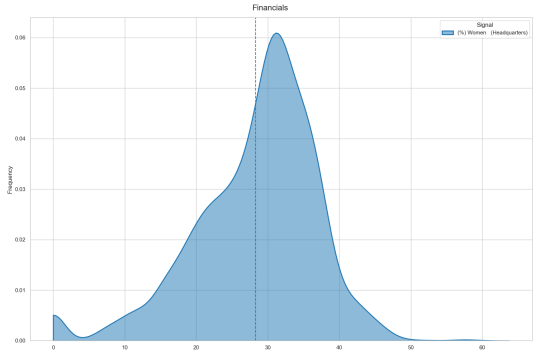
(b) Cyclical sector



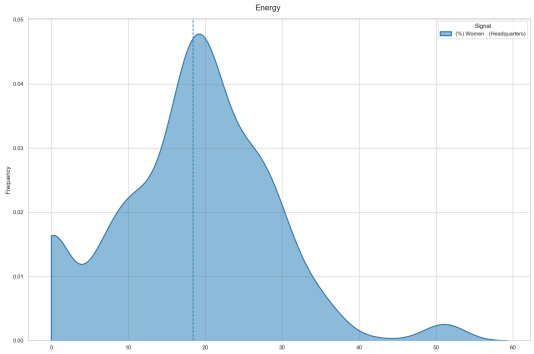
(c) Defensive sector



(d) Growth & Innovation sector



(e) Financials sector



(f) Energy sector

Figure D.1: Kernel density plots of percentage women.

Table D.5: Descriptive Statistics - (Growth & Innovation)

Variable	Min	Mean	Median	Max	Std Dev	Skewness	Kurtosis	N
(%) Women	0.000	25.851	25.806	100	10.846	0.173	2.076	4838
(%) Unknown gender	0.000	0.028	0.023	0.270	0.026	1.973	8.095	4838
Tobin's Q	-0.612	1.225	0.657	5.047	1.636	1.129	0.209	4204
Total assets	10.392	15.646	15.721	20.174	1.845	-0.102	-0.102	4419
Leverage	0.000	0.261	0.244	1.552	0.205	1.029	2.182	4417
Total employees	88.170	38170.464	9814.290	923390.810	85916.556	4.623	26.857	4844

Table D.6: Descriptive Statistics - (Financials)

Variable	Min	Mean	Median	Max	Std Dev	Skewness	Kurtosis	N
(%) Women	0.000	28.311	30.033	58.793	8.580	-0.891	1.354	3521
(%) Unknown gender	0.000	0.040	0.040	0.191	0.025	0.499	1.260	3521
Tobin's Q	-0.612	-0.152	-0.531	5.047	1.031	3.598	13.430	3234
Total assets	11.116	17.715	17.899	22.098	2.121	-0.407	-0.110	3356
Leverage	0.000	0.156	0.091	0.972	0.180	2.113	5.000	3349
Total employees	85.559	27831.534	8832.051	292316.720	49311.604	3.113	9.830	3552

Table D.7: Descriptive Statistics - (Energy)

Variable	Min	Mean	Median	Max	Std Dev	Skewness	Kurtosis	N
(%) Women	0.000	18.430	18.975	51.766	10.192	0.218	0.629	1101
(%) Unknown gender	0.000	0.008	0.000	0.054	0.012	1.355	0.936	1101
Tobin's Q	-0.612	-0.268	-0.327	0.953	0.253	1.448	2.638	1051
Total assets	11.440	16.642	16.702	19.868	1.630	-0.433	0.930	1088
Leverage	0.000	0.303	0.266	0.932	0.174	1.140	1.758	1088
Total employees	238.381	20174.814	3405.016	141472.060	32817.056	1.869	2.576	1120

D.3 Correlation Analysis

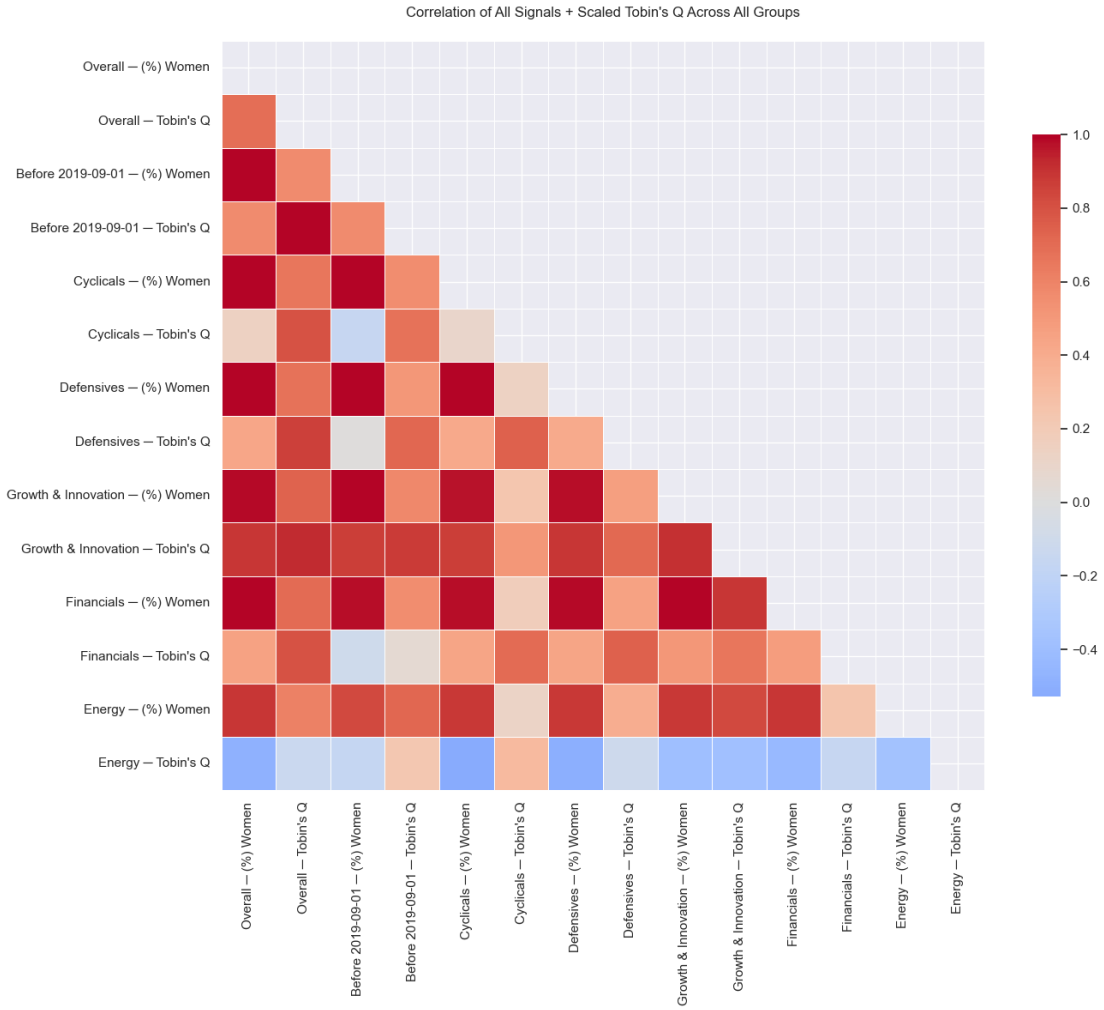
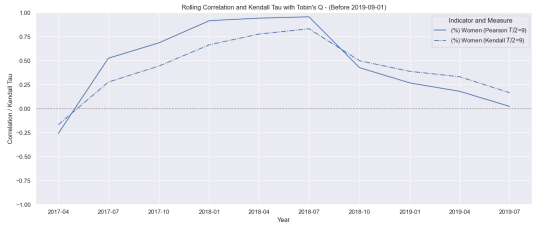
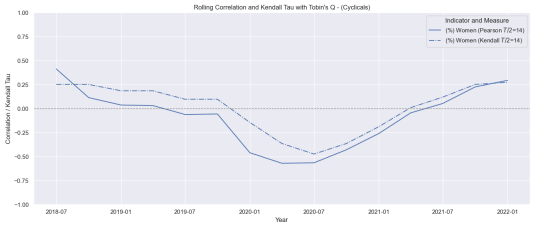


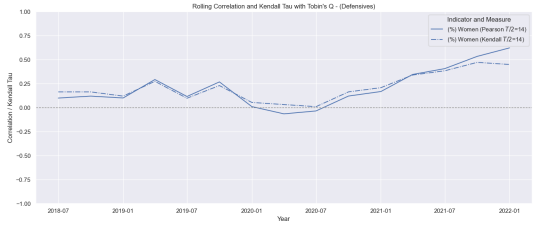
Figure D.2: Pearson correlation heatmap illustrating the relationships between Tobin's Q and gender across all sectoral groups.



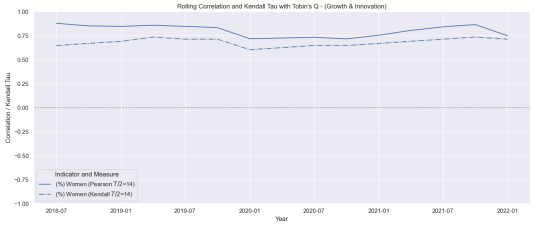
(a) All firms (before 01-09-2019)



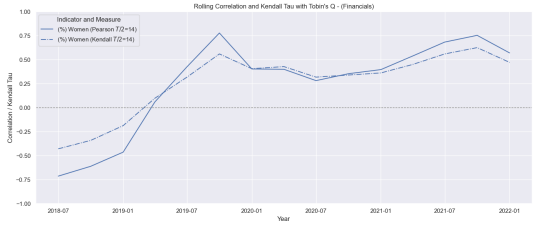
(b) Cyclical sector



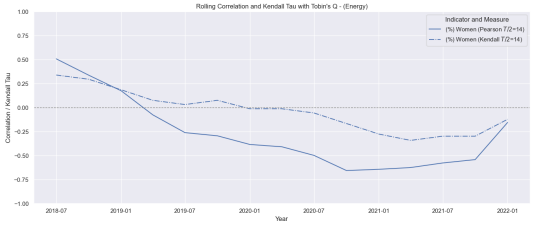
(c) Defensive sector



(d) Growth & Innovation sector



(e) Financials sector



(f) Energy sector

Figure D.3: Rolling correlations plots of gender with Tobin's Q.

D.4 Data Clustering by Threshold

Table D.8: Counts by Threshold and Group – (% Women)

Note: Grey-shaded cells indicate clusters with $N_k < 10$ for $k = 0, 1$.

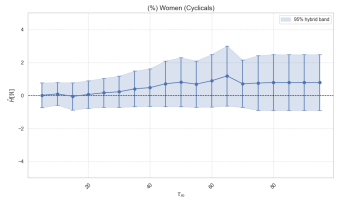
(%) Women	Overall		Pre 01/04/2019		Cyclicals		Defensives		Growth & Innovation		Financials		Energy	
	N_0	N_1	N_0	N_1	N_0	N_1	N_0	N_1	N_0	N_1	N_0	N_1	N_0	N_1
τ_5	1298	23930	1004	15214	734	9822	119	4837	199	4645	110	3446	136	984
τ_{10}	1965	23263	1484	14734	1079	9477	158	4798	317	4527	155	3401	238	882
τ_{15}	3607	21621	2759	13459	2030	8526	246	4710	652	4192	273	3283	368	752
τ_{20}	6886	18342	5004	11214	3674	6882	505	4451	1432	3412	578	2978	635	485
τ_{25}	10902	14326	7685	8533	5538	5018	1102	3854	2272	2572	1068	2488	853	267
τ_{30}	14969	10259	10219	5999	6976	3580	1965	2991	3130	1714	1784	1772	1012	108
τ_{35}	18937	6291	12634	3584	8167	2389	2850	2106	3872	972	2830	726	1074	46
τ_{40}	21853	3375	14332	1886	9027	1529	3708	1248	4441	403	3397	159	1099	21
τ_{45}	23330	1898	15163	1055	9525	1031	4279	677	4703	141	3526	30	1101	19
τ_{50}	24102	1126	15549	669	9855	701	4592	364	4799	45	3552	4	1108	12
τ_{55}	24651	577	15886	332	10113	443	4836	120	4833	11	3553	3	1120	0
τ_{60}	24952	276	16080	138	10335	221	4910	46	4835	9	3556	0	1120	0
τ_{65}	25081	147	16135	83	10445	111	4928	28	4836	8	3556	0	1120	0
τ_{70}	25158	70	16175	43	10491	65	4956	0	4839	5	3556	0	1120	0
τ_{75}	25177	51	16191	27	10510	46	4956	0	4839	5	3556	0	1120	0
τ_{80}	25182	46	16195	23	10515	41	4956	0	4839	5	3556	0	1120	0
τ_{85}	25182	46	16195	23	10515	41	4956	0	4839	5	3556	0	1120	0
τ_{90}	25182	46	16195	23	10515	41	4956	0	4839	5	3556	0	1120	0
τ_{95}	25182	46	16195	23	10515	41	4956	0	4839	5	3556	0	1120	0

Cluster	Diversity Signal	Min	Max
Overall	(%) women	5	95
Pre 01/09/2019	(%) women	5	95
Cyclicals	(%) women	5	95
Defensives	(%) women	5	65
Growth & Innovation	(%) women	5	55
Financials	(%) women	5	45
Energy	(%) women	5	50

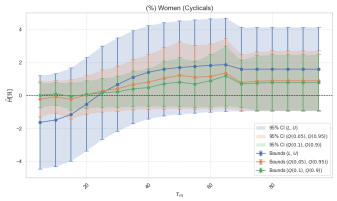
Table D.9: Summary of Minimum and Maximum Permissible Values of τ_m Across Clusters.

Note: The greyed rows are eliminated from the analysis due to small maximum τ_m values, which render them unsuitable for further analysis.

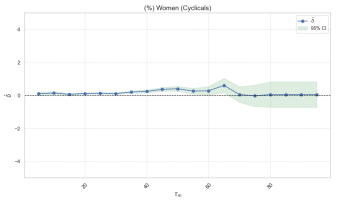
D.5 Partial and Point Identification of the Treatment Effect



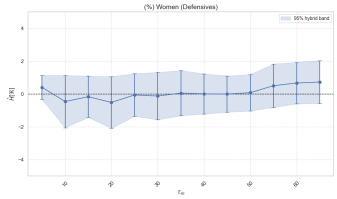
(a) Hybrid: Cyclicals (%) women



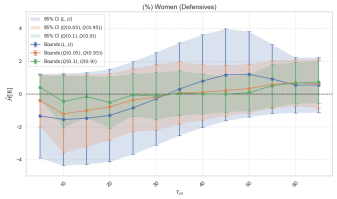
(b) Nonparametric: Cyclicals (%) women



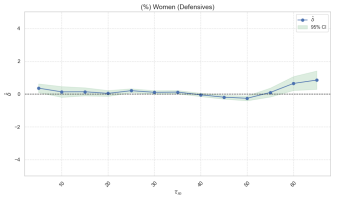
(c) Angrist: Cyclicals (%) women



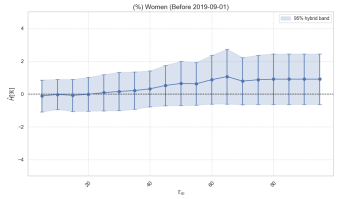
(d) Hybrid: Defensives (%) women



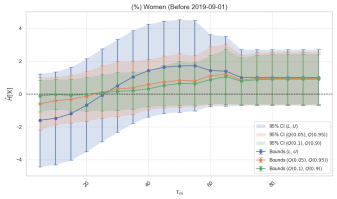
(e) Nonparametric: Defensives (%) women



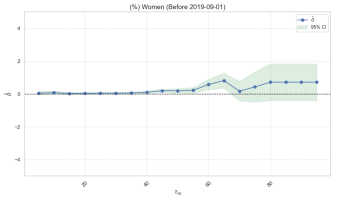
(f) Angrist: Defensives (%) women



(g) Hybrid: Pre 01/04/2019 (%) women

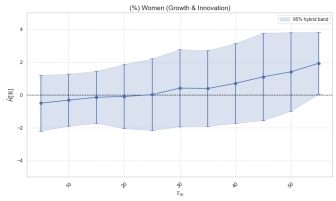


(h) Nonparametric: Pre 01/04/2019 (%) women

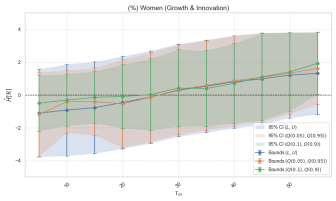


(i) Angrist: Pre 01/04/2019 (%) women

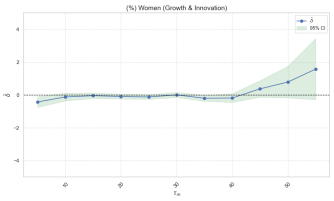
Figure D.4: Hybrid, Manski nonparametric and Angrist estimates



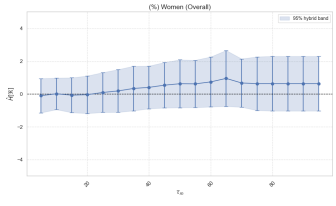
(a) Hybrid: Growth & innovation (%) women



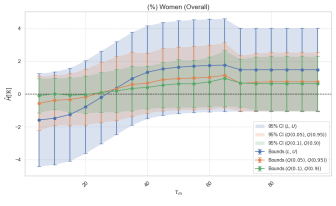
(b) Nonparametric: Growth & innovation (%) women



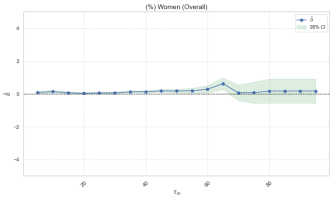
(c) Angrist: Growth & innovation (%) women



(d) Hybrid: Overall (%) women



(e) Nonparametric: Overall (%) women



(f) Angrist: Overall (%) women

Figure D.5: Hybrid, Manski nonparametric and Angrist estimates