



Volume 0, Issue 0

Minimax causal inference under tail heterogeneity

Omid M Ardakani
Georgia Southern University

Abstract

Standard causal estimators deteriorate when treatment modifies the tail index of outcomes. I generalize the maximum domain of attraction to nonstationary triangular arrays, establishing uniform convergence of block maxima to time-varying generalized extreme value distributions. Under such shifts, the supports of the limiting distributions separate, likelihood ratios are unbounded, and inverse probability weighting loses root- n consistency and asymptotic normality once the heavier tail index exceeds one half. I derive minimax lower bounds for the causal estimation rate; a tail-adaptive doubly robust estimator attains them up to logarithms. In simulations, tail adaptation cuts integrated mean absolute error by 20 to 41 percent at crisis-scale samples and stabilizes performance under strong tail heterogeneity. An application to bank stock returns around the 2020 Federal Reserve emergency facilities

I would like to express my gratitude to Wim Vijverberg for his insightful and invaluable feedback. Thank you also to Allen Zinkle, who helped proofread this paper, and to Partha Deb for his advice. Additionally, I acknowledge that the collection of data used in this study was partly supported by the National Institutes of Health under grant number R01 HD069609 and R01 AG040213, and the National Science Foundation under award numbers SES 1157698 and 1623684. Editor's Note: This paper was originally submitted under a different manuscript number on 12/04/2025 and accepted for publication on 07/21/2026. A system error required us to replace the original submission with a new manuscript with a later manuscript number.

Citation: Omid M Ardakani, (2026) "Minimax causal inference under tail heterogeneity", *Economics Bulletin*, Volume 0, Issue 0, pages 1454-1464

Contact: Omid M Ardakani - oardakani@georgiasouthern.edu.

Submitted: October 12, 2026. **Published:** October 12, 2026.

1 Introduction

Extreme events drive financial crises, yet causal inference methods presume well-behaved outcomes (Imbens and Rubin, 2015; Chernozhukov et al., 2017). When an intervention shifts the tail index ξ_t , inverse probability weighting (IPW) breaks down as an inferential tool. Its influence function has infinite variance when $\xi > 0.5$, so $\sqrt{n}$ -consistency and asymptotic normality fail (Embrechts et al., 1997), and the support of generalized extreme value (GEV) distributions shifts when ξ changes. Classical extreme value theory (EVT) assumes stationarity (Fisher and Tippett, 1928; Gnedenko, 1943), while data exhibit drift in tail behavior (Cont, 2001; Hill, 2011; Ardakani, 2023a). The nonstationary EVT literature (Leadbetter et al., 1983; Drees, 2003; Rootzén, 1986) has not addressed causal identification, while the causal inference literature (Rosenbaum and Rubin, 1983; Chernozhukov et al., 2022; Belloni et al., 2017) has not addressed heavy-tailed potential outcomes.

Recent work bridges these fields. Chavez-Demoulin and Mhalla (2024) survey causal inference for extremes, focusing on structural causal models and graph-based approaches. Mhalla et al. (2020) develop an information-theoretic causal discovery method for extreme river discharges. Gnecco et al. (2021) establish causal discovery for heavy-tailed linear models, and Engelke and Ivanovs (2021) develop sparse graphical models for multivariate extremes. Closest to this paper, Deuber et al. (2024) propose a causal Hill estimator for extremal quantile treatment effects with heavy tails. However, all existing work targets either causal structure learning or quantile effects. Minimax optimality for the average treatment effect (ATE) under time-varying tail heterogeneity remains open.

This paper makes three contributions. First, I generalize the Fisher-Tippett-Gnedenko theorem to nonstationary triangular arrays, extending de Haan and Ferreira (2006) and Drees (2003). Second, I show that differing tail indices cause GEV support divergence, yielding infinite Kullback–Leibler (KL) divergence and IPW failure. This extends the mutual singularity results of Resnick (2008) and the information-theoretic framework of Ardakani (2023a, 2025). Third, I derive minimax lower bounds and propose a tail-adaptive doubly robust estimator achieving near-optimal convergence, extending Deuber et al. (2024) from quantile effects to the ATE.

2 Time-varying maximum domain of attraction

Let $\{Y_{t,n}(w) : t = 1, \dots, n\}$ be potential outcomes for $w \in \{0, 1\}$, independent across t but not identically distributed. For blocks $I_n(t)$ with $|I_n(t)| = m_n$, $m_n \rightarrow \infty$, $m_n/\log n \rightarrow \infty$, $m_n = o(n)$. The analysis operates under the following conditions, stated in full here and referenced throughout.

Assumption 1. (a) For each n , the potential outcomes $\{Y_{t,n}(w)\}$ are independent within and across blocks, with row-wise distributions $F_{t,w}$ that may vary with t . (b) Each $F_{t,w}$ has shape parameter $\xi_{t,w} \in [\underline{\xi}, \bar{\xi}] \subset (0, 1)$, and the tail quantile functions $U_{t,w}$ satisfy a second-order regular variation condition uniformly in t . (c) $|\xi_{t,w} - \xi_{s,w}| \leq L|t - s|^\beta$ for some $\beta \in (0, 1]$ and $L < \infty$; the location and scale parameters $\mu_{t,w}$ and $\sigma_{t,w}$ are bounded, with $\sigma_{t,w}$ bounded away from zero. (d) Identification. Strong ignorability $\{Y_t(1), Y_t(0)\} \perp\!\!\!\perp W_t \mid X_t$ and SUTVA hold. (e) $0 < \eta \leq e_t(X_t) \leq 1 - \eta < 1$ almost surely.

Part (b) with $\bar{\xi} < 1$ guarantees $\mathbb{E}|Y_t(w)| < \infty$ (Embrechts et al. 1997, Section 3.4), so the average treatment effect exists; part (b) with $\underline{\xi} > 0$ restricts attention to the heavy-tailed Fréchet domain, the empirically relevant case for financial returns (Cont, 2001).

Definition 1. $\{Y_{t,n}(w)\}$ belongs to the time-varying maximum domain of attraction (MDA) of $\text{GEV}(\mu_{t,w}, \sigma_{t,w}, \xi_{t,w})$ if there exist $a_{t,w}(n) > 0$, $b_{t,w}(n) \in \mathbb{R}$ such that

$$\sup_{t \in [T_n]} \left| \mathbb{P} \left(\frac{\max_{i \in I_n(t)} Y_{i,n}(w) - b_{t,w}(n)}{a_{t,w}(n)} \leq y \right) - G_{t,w}(y) \right| \rightarrow 0, \quad (1)$$

where $G_{t,w}(y) = \exp(-(1 + \xi_{t,w}y)^{-1/\xi_{t,w}})$ for $1 + \xi_{t,w}y > 0$, and $T_n = \lfloor n/m_n \rfloor$.

Theorem 1. Under Assumption 1(a)–(c), the following are equivalent: (i) (1) holds; (ii) the tail quantile function $U_{t,w}(s) = \inf\{y : F_{t,w}(y) \geq 1 - 1/s\}$ satisfies, uniformly in t ,

$$\lim_{s \rightarrow \infty} \sup_{y \geq y_0} \left| \frac{U_{t,w}(sy) - U_{t,w}(s)}{a_{t,w}(s)} - \frac{y^{\xi_{t,w}} - 1}{\xi_{t,w}} \right| = 0 \quad \forall y_0 > 0;$$

(iii) the conditional excess distribution converges uniformly to the generalized Pareto distribution $\text{GPD}(\xi_{t,w})$.

Proof. (i) $\Rightarrow$ (ii): For fixed t , de Haan and Ferreira (2006) gives $U_{t,w}(sy)/a_{t,w}(s) \rightarrow (y^{\xi_{t,w}} - 1)/\xi_{t,w}$. Uniformity in t follows because the block structure and the Hölder condition on $\xi_{t,w}$ ensure the regularly varying functions $U_{t,w}$ form an equicontinuous family; see Drees (2003) and de Haan and Ferreira (2006) for uniform convergence of regularly varying functions under parameter continuity. (ii) $\Rightarrow$ (iii): With $U_{t,w}(sy) \approx U_{t,w}(s) + a_{t,w}(s)(y^{\xi_{t,w}} - 1)/\xi_{t,w}$ uniformly, setting $s = n$ gives

$$\mathbb{P}(Y_{t,n}(w) > U_{t,w}(n) + ya_{t,w}(n) \mid Y_{t,n}(w) > U_{t,w}(n)) \rightarrow (1 + \xi_{t,w}y)^{-1/\xi_{t,w}}$$

uniformly in t and $y > y_0$, which is the GPD with shape $\xi_{t,w}$. (iii) $\Rightarrow$ (i): The Balkema-de Haan theorem (Embrechts et al., 1997, Section 3.4) gives pointwise MDA membership for each t . By de Haan and Ferreira (2006, Theorem 1.1.6), normalizing sequences satisfy $b_{t,w}(n) = U_{t,w}(n) + o(a_{t,w}(n))$, $\tilde{a}_{t,w}(n) \sim a_{t,w}(n)$. Uniformity follows because the excess convergence in (iii) is uniform over t , ensuring the normalizations are uniformly valid.

Remark 1. Independence within the array, Assumption 1(a), is the substantive restriction for financial data; two observations temper it. First, raw returns exhibit volatility clustering; the framework then applies to GARCH-filtered residuals, in the spirit of McNeil and Frey (2000), which restores approximate independence and preserves the conditional tail index. Second, the block construction in Definition 1 extends to weakly dependent sequences satisfying the $D(u_n)$ and $D'(u_n)$ conditions of Leadbetter et al. (1983). Block maxima obey the same GEV limits with an extremal-index adjustment to the effective block size, and the uniformity argument in Theorem 1 carries over because the Hölder condition on $\xi_{t,w}$ is unaffected. A full treatment of the dependent case is beyond this note. Hölder continuity itself is mild. It accommodates smooth drift in tail thickness and approximates regime shifts arbitrarily well through steep transitions, at the cost of a larger L .

Remark 2. The framework handles distributional nonstationarity, drift in $(\mu_{t,w}, \sigma_{t,w}, \xi_{t,w})$ across rows of the array, in the sense of locally stationary processes (Dahlhaus, 1997). Definition 1 imposes no stationarity across t ; uniform convergence of normalized block maxima replaces the fixed limit of the Fisher–Tippett–Gnedenko theorem. Excluded are explosive dynamics in the outcome level and discontinuous breaks in $\xi_{t,w}$, though the latter are approximable per Remark 1.

Under Assumption 1(d)–(e), the ATE $\theta_t = \mathbb{E}[Y_t(1) - Y_t(0)]$ is identified (Rosenbaum and Rubin, 1983), and Assumption 1(b) guarantees it is finite.

2.1 Singularity and minimax bounds

Lemma 1. Let $F_{t,w} = \text{GEV}(\mu_{t,w}, \sigma_{t,w}, \xi_{t,w})$ with $\xi_{t,w} > 0$. The lower support bound is $L_{t,w} = \mu_{t,w} - \sigma_{t,w}/\xi_{t,w}$. If $L_{t,1} > L_{t,0}$, then $\mathcal{D}_{\text{KL}}(F_{t,0}; F_{t,1}) = \infty$.

Proof. The density $f_{t,1}(y) = 0$ for $y < L_{t,1}$, while $f_{t,0}(y) > 0$ for $y \in (L_{t,0}, L_{t,1})$ since $L_{t,0} < L_{t,1}$. Thus $\mathcal{D}_{\text{KL}}(F_{t,0}; F_{t,1}) = \int f_{t,0} \log(f_{t,0}/f_{t,1}) dy \geq \int_{L_{t,0}}^{L_{t,1}} f_{t,0} \log(f_{t,0}/0) dy = +\infty$. When

$\mu_{t,0} = \mu_{t,1}$, $\sigma_{t,0} = \sigma_{t,1} = \sigma$, and $\xi_{t,1} > \xi_{t,0} > 0$, we have $L_{t,1} = \mu - \sigma/\xi_{t,1} > \mu - \sigma/\xi_{t,0} = L_{t,0}$, since $1/\xi_{t,1} < 1/\xi_{t,0}$. This holds generically when treatment increases ξ .

Lemma 1 has two consequences whose scope deserves precise statement. First, infinite KL divergence means the treated and control extreme value laws are mutually singular on the region below $L_{t,1}$, extending the singularity results of Resnick (2008) to causal settings: no common dominating structure supports likelihood-ratio arguments, and the semiparametric efficiency theory of van der Vaart (2000) does not apply uniformly over the model class. Second, the failure of IPW under tail heterogeneity is a failure of rates and inference rather than of consistency per se. Under Assumption 1, $\bar{\xi} < 1$ ensures the ATE exists and IPW remains consistent; however, once $\max_w \xi_{t,w} > 1/2$, the influence function has infinite variance, so $\sqrt{n_t}$ -consistency and asymptotic normality fail and the convergence rate degrades to $n_t^{-(1-\max_w \xi_{t,w})}$ up to slowly varying factors (Embrechts et al., 1997). This rate degradation, combined with the loss of a valid normal approximation for inference, is what the simulations in Section 2.3 quantify, and it provides theoretical grounding for the instability observed by Bollerslev et al. (2020) and Ardakani (2023b).

Theorem 2. *Let $\mathcal{P}$ be the class of distributions in the time-varying MDA of Definition 1 satisfying Assumption 1 with parameters $(\underline{\xi}, \bar{\xi}, L, \beta)$. Then*

$$\inf_{\hat{\theta}_t} \sup_{P \in \mathcal{P}} \mathbb{E}_P |\hat{\theta}_t - \theta_t| \geq c n_t^{-\beta/(2\beta+1)}, \quad (2)$$

where $c > 0$ depends on $\underline{\xi}, \bar{\xi}, L, \beta$.

Proof. Construct hypotheses $P_0: \xi_{t,w}^{(0)} = \xi_0$ and $P_1: \xi_{t,w}^{(1)} = \xi_0 + \delta_n \phi((t - t_0)/h_n)$, where $\phi \in C^\beta(\mathbb{R})$ with $\phi(0) = 1$, $\text{supp}(\phi) \subset [-1, 1]$, $\|\phi\|_{C^\beta} \leq 1$. Set $\delta_n = c_1 n_t^{-\beta/(2\beta+1)}$ and $h_n = c_2 n_t^{-1/(2\beta+1)}$. The Hölder constraint $\delta_n \leq L h_n^\beta$ is satisfied by construction. Under P_1 , $\theta_{t_0} = \theta_0 + c_3 \delta_n$ where $c_3 = \sigma \partial_\xi[(\Gamma(1 - \xi) - 1)/\xi]|_{\xi=\xi_0}$, bounded away from zero on $[\underline{\xi}, \bar{\xi}]$.

The perturbation affects $\sim 2h_n$ periods around t_0 . At each affected period, n_t observations have Hellinger distance $H^2(p_{0,t}, p_{1,t}) \leq C \delta_n^2 \phi^2((t - t_0)/h_n)$ by a quadratic Hellinger bound for the GEV family; the parameter-dependent lower endpoint contributes only exponentially small mass because the GEV density vanishes to all orders there (Tsybakov, 2009). The product Hellinger distance satisfies

$$H^2(P_0^\otimes, P_1^\otimes) \leq \sum_{|t-t_0| \leq h_n} n_t H^2(p_{0,t}, p_{1,t}) \leq c_4 n_t h_n \delta_n^2.$$

With $\delta_n \asymp n_t^{-\beta/(2\beta+1)}$ and $h_n \asymp n_t^{-1/(2\beta+1)}$, we obtain $n_t h_n \delta_n^2 = n_t \cdot n_t^{-1/(2\beta+1)} \cdot n_t^{-2\beta/(2\beta+1)} = 1$. Le Cam's lemma (Tsybakov, 2009, Theorem 2.2) gives the lower bound.

The rate $n_t^{-\beta/(2\beta+1)}$ matches Stone's nonparametric lower bound (Stone, 1980) and Drees (1998)'s optimal rate for tail index estimation, confirming that the bottleneck is estimating time-varying $\xi_{t,w}$.

2.2 Tail-adaptive doubly robust estimator

The estimator proceeds in three steps: estimate the tail index, convert it into observation weights, and average a weighted doubly robust score.

Step 1: tail-index estimation. For each (t, w) , let $N_{t,w}$ denote the number of arm- w observations at period t and let $Y_{(1)} \leq \dots \leq Y_{(N_{t,w})}$ be their order statistics. The tail index is estimated by the Hill estimator (Hill, 1975),

$$\hat{\xi}_{t,w} = \frac{1}{k_{t,w}} \sum_{j=1}^{k_{t,w}} \log \frac{Y_{(N_{t,w}-j+1)}}{Y_{(N_{t,w}-k_{t,w})}}, \quad k_{t,w} = \lfloor N_{t,w}^{2\beta/(2\beta+1)} \rfloor, \quad (3)$$

with threshold $\hat{b}_{t,w} = Y_{(N_{t,w}-k_{t,w})}$ and scale $\hat{\sigma}_{t,w} = (1 - \hat{\xi}_{t,w}) \bar{e}_{t,w}$, where $\bar{e}_{t,w}$ is the mean excess over $\hat{b}_{t,w}$; this is the moment estimator implied by the generalized Pareto mean excess function. When n_t is small, observations are pooled over a local window $[t - h_n, t + h_n]$ with $h_n \asymp n_t^{-1/(2\beta+1)}$ in rescaled time, which the Hölder condition in Assumption 1(c) permits at the cost of a drift bias of order h_n^β . The choice of $k_{t,w}$ balances the Hill variance $k_{t,w}^{-1/2}$ against this bias, so $|\hat{\xi}_{t,w} - \xi_{t,w}| = O_P(k_{t,w}^{-1/2} + h_n^\beta) = O_P(n_t^{-\beta/(2\beta+1)})$ (Drees, 1998).

Step 2: tail-adaptive weights. Define $\omega_i = \min\{1, (1 + \hat{\xi}_{t,w_i}(Y_i - \hat{b}_{t,w_i})/\hat{\sigma}_{t,w_i})_+^{-1}\}$: observations below the threshold receive unit weight, and exceedances are downweighted along the fitted generalized Pareto survival curve of their own treatment arm.

Step 3: weighted doubly robust average. With ψ_i the standard augmented score, the estimator is the normalized weighted average

$$\hat{\theta}_t = \frac{\sum_{i:T_i=t} \omega_i \psi_i}{\sum_{i:T_i=t} \omega_i}, \quad \psi_i = \hat{\mu}_{t,1}(X_i) - \hat{\mu}_{t,0}(X_i) + \frac{W_i(Y_i - \hat{\mu}_{t,1}(X_i))}{\hat{e}_t(X_i)} - \frac{(1 - W_i)(Y_i - \hat{\mu}_{t,0}(X_i))}{1 - \hat{e}_t(X_i)}. \quad (4)$$

The mechanism through which the weights improve estimation under heavy tails is variance restoration. As $Y_i \rightarrow \infty$, $\omega_i Y_i \rightarrow \hat{\sigma}_{t,w}/\hat{\xi}_{t,w}$, so every weighted contribution to (4) is bounded even when $\text{Var}(Y_t(w)) = \infty$, that is, when $\xi_{t,w} \geq 1/2$ and the unweighted score has no second moment. The price is a smoothing bias of order $n_t^{-\beta/(2\beta+1)}$ inherited from $\hat{\xi}_{t,w}$, which does not vanish faster than the tail-estimation rate. Combining this with the doubly robust remainder $O_P(\|\hat{\mu} - \mu\| \cdot \|\hat{e} - e\|)$ (Chernozhukov et al., 2017) yields $\mathbb{E}|\hat{\theta}_t - \theta_t| \leq C n_t^{-\beta/(2\beta+1)} \log n_t$, matching the lower bound in Theorem 2 up to the logarithmic factor. The bias-variance structure predicts exactly when tail adaptation pays: when tails are heavy enough that the variance term dominates, that is, in small and moderate samples, and it predicts when it does not: in large samples with moderate heterogeneity, where the bias floor binds while the competitors' variance has already shrunk. The simulations confirm both predictions.

2.3 Simulation

For $t = 1, \dots, T$ and $i = 1, \dots, n_t$: $X_{it} \sim N(0, 1)$; $e_t(x) = \Lambda(0.4 \sin(2\pi t/T) + 0.5x)$, Λ logistic; $Y_{it}(w) = w + X_{it} + Z_{it,w}$ with $Z_{it,w} \sim \text{GEV}(0, 1, \xi_{t,w})$ and $\xi_{t,w} = \xi_w + 0.05 \sin(2\pi t/T)$, so propensity and tail index drift smoothly (Assumption 1). The true effect $\theta_t = 1 + m(\xi_{t,1}) - m(\xi_{t,0})$, with $m(\xi) = (\Gamma(1 - \xi) - 1)/\xi$ the GEV mean, is closed form. Base

indices define five scenarios: homogeneous ($\xi_0 = \xi_1 = 0.4$), heterogeneous (0.3, 0.6), strong (0.2, 0.8), light (0.1, 0.1), and misspecified (heterogeneous tails, intercept-only propensity). Per period, nuisances are a logit of W on $(1, X)$ for $\hat{e}_t$ and per-arm least squares of Y on $(1, X)$ for $\hat{\mu}_{t,w}$. IPW is the normalized (Hájek) estimator; trimmed IPW drops Y outside the period’s 5th–95th percentiles before reweighting, an analogue of systematic trimming (Crump et al., 2009); DR sets $\omega_i \equiv 1$; the tail-adaptive estimator implements (3)–(4) with $\beta = 1$; the oracle uses true nuisances. Performance is integrated MAE, $T^{-1} \sum_t |\hat{\theta}_t - \theta_t|$, over 500 replications, scaled by 100; Monte Carlo standard errors are across-replication standard deviations over $\sqrt{500}$. Replication code in Python and R accompanies the paper. Table I varies n_t at $T = 50$; Table II varies T at $n_t = 200$.

Table I: Integrated MAE ($\times 100$) across sample sizes, 500 replications

n_t	Scenario	IPW	Trim. IPW	DR	Tail-Ad. DR	Oracle	Max MC se
100	Homogeneous	54.3	34.5	54.2	32.0	53.5	0.38
	Heterogeneous	95.9	122.0	96.1	76.8	94.9	1.88
	Strong hetero.	358.6	384.1	360.9	346.1	355.0	18.82
200	Homogeneous	39.6	29.4	39.5	25.1	39.2	0.29
	Heterogeneous	74.4	123.9	74.5	73.0	74.0	1.43
	Misspecified	74.3	88.3	74.4	78.6	74.0	1.48
500	Homogeneous	26.1	26.5	26.0	17.6	25.9	0.17
	Heterogeneous	52.2	124.5	52.2	69.7	52.1	0.73
	Light tail	11.3	28.5	11.3	12.3	11.2	0.07
1000	Homogeneous	18.7	26.1	18.6	13.4	18.6	0.11
	Heterogeneous	40.1	124.9	40.1	67.2	40.1	0.50
	Strong hetero.	211.0	390.5	211.1	318.3	211.5	5.86

Bold: lowest MAE, oracle (true nuisances) excluded. Trimmed IPW drops Y outside 5th–95th percentiles. (ξ_0, ξ_1) : homogeneous (0.4, 0.4); heterogeneous (0.3, 0.6); strong (0.2, 0.8); light (0.1, 0.1); misspecified: (0.3, 0.6) with intercept-only propensity. Max MC se: largest Monte Carlo standard error in the row.

Table I reveals a bias-variance tradeoff that follows the theory of Section 2.2. In small and moderate samples the tail-adaptive estimator dominates wherever tails are heavy. At $n_t = 100$ it cuts integrated MAE relative to DR by 20% under heterogeneous tails, 41% under homogeneous heavy tails, and 4% under strong heterogeneity; the pattern persists at $n_t = 200$. As n_t grows, competitors’ variance shrinks at rate $n_t^{-1/2}$ where second moments exist, while the tail-adaptive smoothing bias of order $n_t^{-\beta/(2\beta+1)}$ binds. The log-log slope of MAE against n_t under homogeneity is -0.46 for DR, near the parametric -0.5 , against -0.38 for the tail-adaptive estimator, near the minimax exponent $-1/3$ of Theorem 2 with $\beta = 1$. DR therefore overtakes under moderate heterogeneity at $n_t \geq 500$, the crossover the bias floor predicts. Under strong heterogeneity, where $\xi_{t,1}$ exceeds the infinite-variance threshold, the comparison shifts from accuracy to stability. The tail-adaptive Monte Carlo standard error never exceeds 2.3, while the standard estimators’ reach 5.9 in Table I and 50.0 in Table II; their realized-error distribution is itself heavy-tailed, so their average MAE is poorly pinned down even with 500 replications. Trimmed IPW competes only under homogeneous tails in small samples and fails under heterogeneity, with a bias near 124 that does not shrink in

n_t , because symmetric trimming discards informative extremes asymmetrically across arms. Under the misspecified propensity, the doubly robust estimators retain accuracy through the correct outcome model.

Table II addresses whether the fixed $T = 50$ drives these conclusions. It does not: holding $n_t = 200$ and varying T from 50 to 1000 leaves every estimator’s integrated MAE essentially unchanged under heterogeneous tails, confirming that accuracy is governed by the cross-sectional size n_t , as the theory implies, since θ_t is estimated period by period and the integrated MAE averages over periods. The one visible effect of larger T is distributional: under strong heterogeneity the standard estimators’ realized MAE fluctuates across designs (315 to 347) with Monte Carlo standard errors up to 50, while the tail-adaptive estimator is pinned within one point of 338 throughout.

Table II: Robustness to the number of periods: integrated MAE ($\times 100$) with Monte Carlo standard errors, $n_t = 200$, 500 replications

T	Scenario	IPW	Trim. IPW	DR	Tail-Ad. DR	Oracle
50	Heterogeneous	74.4 (1.43)	123.9 (0.17)	74.5 (1.43)	73.0 (0.21)	74.0 (1.42)
	Strong hetero.	315.1 (18.84)	387.9 (0.20)	315.5 (18.86)	338.5 (1.57)	314.4 (18.70)
100	Heterogeneous	75.1 (1.40)	123.8 (0.12)	75.1 (1.39)	73.2 (0.15)	74.3 (1.26)
	Strong hetero.	326.0 (23.06)	388.0 (0.14)	326.2 (22.80)	338.8 (1.67)	319.9 (20.31)
500	Heterogeneous	74.0 (1.21)	123.7 (0.05)	74.1 (1.22)	73.1 (0.07)	73.7 (1.20)
	Strong hetero.	346.2 (49.57)	387.8 (0.06)	347.0 (49.98)	338.9 (1.62)	345.2 (48.98)
1000	Heterogeneous	74.5 (0.67)	123.8 (0.04)	74.5 (0.67)	73.2 (0.04)	74.1 (0.62)
	Strong hetero.	335.0 (19.06)	388.0 (0.04)	335.5 (19.11)	338.0 (0.41)	332.5 (17.36)

MC standard errors in parentheses; bold is lowest MAE among estimators; scenarios as in Table I.

The practical conditions under which tail adaptation is most valuable follow directly. The advantages are largest when the sample per period is small ($n_t \leq 200$), tails are heavy in at least one arm, and stability of the estimate across realizations matters, which is precisely the configuration of financial crisis applications: short episode windows, tail indices near or above one half, and decisions that cannot tolerate estimator breakdown in the realization actually observed. When samples are large and tails are moderate or light, standard DR is preferable, and light-tail row of Table I shows the cost of unnecessary adaptation is mild.

3 Empirical application

I apply the framework to daily log-returns of 16 bank stocks around the March 2020 Federal Reserve emergency facilities with macro indicators from FRED. Eight G-SIBs (JPM, BAC, C, WFC, GS, MS, BK, STT) with direct facility access form the treatment group; eight regional banks (ZION, HBAN, CFG, KEY, RF, FHN, FITB, EWBC) serve as controls. Two caveats frame the exercise. G-SIB designation is not randomly assigned; it is bound up with size, business mix, and interconnectedness, and the eight-versus-eight panel is small. The application is therefore an illustration of how time-varying tail heterogeneity manifests in the data and interacts with standard causal estimators, not strong causal evidence on the effects of the facilities themselves.

Table III: Summary statistics and tail analysis of bank stock returns

Period	Group	Return characteristics				Tail and information			
		Ann.Ret	Ann.Vol	VaR _{1%}	Kurt	$\hat{\xi}_H$	(se)	$\hat{H}(r)$	KL
Pre-crisis	G-SIB	0.01	0.25	-4.43	2.1	0.334	0.073	-2.96	0.08
	Regional	-0.15	0.27	-4.83	1.4	0.333	0.073	-2.84	
Crisis	G-SIB	-0.38	0.87	-15.47	1.9	0.400	0.107	-1.84	0.57
	Regional	-0.59	1.04	-17.48	0.6	0.304	0.081	-1.63	
Recovery	G-SIB	0.46	0.35	-4.25	4.4	0.321	0.076	-2.64	0.25
	Regional	0.67	0.48	-5.21	6.1	0.320	0.076	-2.35	
Post-crisis	G-SIB	0.31	0.27	-3.97	0.9	0.256	0.049	-2.78	0.10
	Regional	0.36	0.32	-4.82	1.3	0.246	0.047	-2.59	

VaR_{1%} in percentage points; $\hat{H}(r)$ is Shannon entropy of daily returns; KL is the Kullback-Leibler divergence between pooled G-SIB and Regional returns. Eight stocks per group, 756 trading days; period-average VIX 15.4, 40.5, 25.7, 19.7.

Tables III and IV yield four findings. First, the KL divergence between G-SIB and Regional return distributions increases sevenfold during the crisis, consistent with the distributional divergence mechanism of Lemma 1. Shannon entropy drops sharply, reflecting information compression during extreme stress (Ardakani, 2023a; Ardakani et al., 2025).

Second, G-SIBs exhibit significantly heavier tails than Regional banks during the crisis ($\hat{\xi} = 0.400$ vs. 0.304 , $t = 2.68$, MW $p = 0.028$). This suggests that systemically important banks, being more interconnected (Acharya et al., 2017), absorbed fat-tailed systemic events that regional banks partially avoided. The DR estimate of the ATE on ξ itself is $+0.08$, suggesting that Fed-connected banks experienced heavier tail risk.

Third, the tail-adaptive correction to the volatility ATE is modest (0.8% from DR) because realized $\hat{\xi}$ remains below 0.5. This is consistent with the simulation: the estimator's advantage concentrates when $\xi > 0.5$ and the GEV variance diverges. The crisis $\hat{\xi}$ values approach but do not cross this threshold.

Fourth, outside the crisis, tail heterogeneity is insignificant ($p \geq 0.88$), KL returns to baseline, and estimators agree. This is consistent with the time-varying MDA framework: the tail divergence mechanism is crisis-specific.

Table IV: Tail heterogeneity tests and treatment effect estimates

Period	ξ -heterogeneity			ATE on volatility				ATE on	
	$\Delta\hat{\xi}$	t	MW- p	Naive	IPW	DR	Tail-Ad.	ξ	$H(r)$
Pre-crisis	0.00	0.03	0.878	-0.03	-0.04	-0.01	-0.01	-0.02	-0.07
Crisis	0.10	2.68	0.028	-0.17	-0.08	-0.17	-0.17	0.08	-0.18
Recovery	0.00	0.02	0.959	-0.14	-0.08	-0.14	-0.14	-0.01	-0.30
Post-crisis	0.01	0.49	0.878	-0.06	-0.03	-0.06	-0.06	0.01	-0.21

$\Delta\hat{\xi} = \hat{\xi}_{\text{G-SIB}} - \hat{\xi}_{\text{Regional}}$ (Hill); MW- p : Mann-Whitney test. Outcome models are arm means; the propensity uses excess kurtosis and skewness, plus entropy for the volatility ATE. The ξ and $H(r)$ ATEs use DR.

4 Conclusion

This paper establishes minimax-optimal causal inference under time-varying tail heterogeneity. The time-varying MDA framework generalizes the Fisher-Tippett-Gnedenko theorem to nonstationary potential outcomes. Differing tail indices across treatment arms create support divergence and infinite KL divergence, undermining root- n inference for standard estimators. The tail-adaptive doubly robust estimator achieves near-optimal convergence rates. Empirically, the KL divergence between G-SIB and Regional bank returns increases sevenfold during the 2020 crisis, with G-SIBs exhibiting significantly heavier tails. Extensions to infinite-mean distributions (Gronwald, 2019), long-memory extremes (Zhang, 2021), and high-dimensional covariates via cross-fitting (Chernozhukov et al., 2017; Belloni et al., 2017) remain open.

References

- Acharya, V. V., L. H. Pedersen, T. Philippon, and M. Richardson (2017). Measuring systemic risk. *Review of Financial Studies* 30(1), 2–47.
- Ardakani, O. M. (2023a). Capturing information in extreme events. *Economics Letters* 231, 111301.
- Ardakani, O. M. (2023b). Coherent measure of portfolio risk. *Finance Research Letters* 57, 104222.
- Ardakani, O. M. (2025). Strategic information asymmetry in tail-risk markets. *North American Journal of Economics and Finance* 79, 102460.
- Ardakani, O. M., V. Dalko, and H. Shim (2025). Information loss from perception alignment. *International Review of Economics and Finance* 97, 103830.
- Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* 85(1), 233–298.
- Bollerslev, T., S. Z. Li, and B. Zhao (2020). Good volatility, bad volatility, and the cross section of stock returns. *Journal of Financial and Quantitative Analysis* 55(3), 751–781.
- Chavez-Demoulin, V. and L. Mhalla (2024). Causality and extremes. *arXiv preprint arXiv:2403.05331*.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review* 107(5), 261–265.
- Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022). Locally robust semiparametric estimation. *Econometrica* 90(4), 1501–1535.
- Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance* 1(2), 223–236.

- Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96(1), 187–199.
- Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics* 25(1), 1–37.
- de Haan, L. and A. Ferreira (2006). *Extreme value theory: An introduction*. Springer.
- Deuber, D., J. Li, S. Engelke, and M. H. Maathuis (2024). Estimation and inference of extremal quantile treatment effects for heavy-tailed distributions. *Journal of the American Statistical Association* 119(547), 2206–2216.
- Drees, H. (1998). Optimal rates of convergence for estimates of the extreme value index. *Annals of Statistics* 26(1), 434–448.
- Drees, H. (2003). Extreme quantile estimation for dependent data, with applications to finance. *Bernoulli* 9(4), 617–657.
- Embrechts, P., C. Klüppelberg, and T. Mikosch (1997). *Modelling extremal events for insurance and finance*. Springer.
- Engelke, S. and J. Ivanovs (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application* 8, 241–270.
- Fisher, R. A. and L. H. C. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 24, pp. 180–190.
- Gnecco, N., N. Meinshausen, J. Peters, and S. Engelke (2021). Causal discovery in heavy-tailed models. *Annals of Statistics* 49(3), 1755–1778.
- Gnedenko, B. (1943). Sur la distribution limite du terme maximum d’une serie aleatoire. *Annals of Mathematics* 44(3), 423–453.
- Gronwald, M. (2019). Is Bitcoin a commodity? On price jumps, demand shocks, and certainty of supply. *Journal of International Money and Finance* 97, 86–92.
- Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics* 3(5), 1163–1174.
- Hill, J. B. (2011). Tail and nontail memory with applications to extreme value and robust statistics. *Econometric Theory* 27(4), 844–884.
- Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983). *Extremes and related properties of random sequences and processes*. Springer.

- McNeil, A. J. and R. Frey (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance* 7(3–4), 271–300.
- Mhalla, L., V. Chavez-Demoulin, and D. J. Dupuis (2020). Causal mechanism of extreme river discharges in the upper Danube basin network. *Journal of the Royal Statistical Society: Series C* 69(4), 741–764.
- Resnick, S. I. (2008). *Extreme values, regular variation, and point processes*. Springer.
- Rootzén, H. (1986). Extreme value theory for moving average processes. *Annals of Probability* 14(2), 612–652.
- Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55.
- Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics* 8(6), 1348–1360.
- Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer.
- van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.
- Zhang, T. (2021). High-quantile regression for tail-dependent time series. *Biometrika* 108(1), 113–126.