Applying Informal Benchmarking to the f-Sensitivity Model

Benchmarking the Unobserved

Adrians Slics
Supervisor(s): Jesse Krijthe, Matej Havelka
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Adrians Slics
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Matej Havelka, Avishek Anand

Abstract

Sensitivity analysis asks how much unobserved confounding would overturn a causal conclusion. Every framework leaves the analyst to choose how much confounding to allow for. For the marginal sensitivity model (MSM), *informal benchmarking* sets this choice from the data. Each observed covariate is dropped in turn, and the resulting shift in treatment odds is taken as a plausible value. We ask whether the same idea transfers to the f-sensitivity model, whose parameter ρ bounds confounding by an average within each covariate value rather than by a single worst case. We show that it does. The transfer relies on a single new quantity, a benchmark $\hat{\rho}_{\text{bench}}$. This is the symmetric-KL divergence that a dropped covariate induces between the treatment arms. We take the strongest covariate rather than the average, as informal benchmarking does for the MSM and as ρ requires. We compute $\hat{\rho}_{\text{bench}}$ from the covariates. It is stable across seeds, and it separates covariates that the MSM treats as identical. As a rare confounding spike grows, $\hat{\rho}_{\text{bench}}$ stays nearly flat while the MSM's worst-case reading climbs, which behavior is to be expected of. On simulated data with a known hidden confounder, the benchmark recovers the divergence that the confounder induces, and it covers the true ρ in every scenario tested. It has its shortcomings as it can under-report confounding that is concentrated in a low-density region of a covariate's range.

1 Introduction

Causal inference from observational data usually assumes that all confounders are observed (Yao et al., 2021). This assumption rarely holds. Sensitivity analysis takes a different approach. It measures how unobserved confounding could change a conclusion. Several frameworks do this. They include the marginal sensitivity model (Tan, 2006), the f-sensitivity model (Jin et al., 2022), and the omitted-variable-bias framework of Cinelli and Hazlett (2020) and Chernozhukov et al. (2024). Each one asks the analyst to set a parameter that bounds the strength of hidden confounding. Clear guidance for that choice is often missing.

Informal benchmarking gives a partial answer. It uses the observed covariates as reference points for plausible parameter values. The idea goes back to Imbens (2003). Baitairian et al. (2025) review it for the marginal sensitivity model. It has not been studied for the f-sensitivity model (Jin et al., 2022). This report addresses that gap.

Can the idea of informal benchmarking be applied to the f-sensitivity model, and if so, does it yield meaningful sensitivity parameter estimates?

We split the question into five sub-questions, namely (i) how is informal benchmarking defined in the marginal sensitivity model and what are its formal assumptions; (ii) what structural differences between the two models affect the transferability of the procedure; (iii) what does a valid adaptation look like; (iv) under what conditions does the adapted procedure produce accurate parameter estimates; and (v) under what conditions does it fail.

The report makes three contributions. First, it analyses the structural differences between the two models that affect the transfer. Second, it builds the adaptation around a divergence benchmark that measures each covariate by its average shift between the treatment arms rather than by its worst value. Third, it evaluates that benchmark in simulation, where it is stable across seeds and behaves as an average-divergence budget should, staying flat under a rare confounding spike while the MSM's worst-case reading climbs, and recovering the divergence a known simulated confounder induces (covering it on both scales), while under-reporting confounding concentrated in a low-density region of a covariate's range.

2 Background

2.1 Causal inference

In the potential-outcomes framework (Yao et al., 2021), each unit has a binary treatment $T \in \{0,1\}$, observed covariates X, and potential outcomes $Y(0), Y(1)$. We observe only $Y = T\,Y(1) + (1-T)\,Y(0)$. The estimand is the average treatment effect (ATE) $\tau = \mathbb{E}[Y(1) - Y(0)]$. This is the outcome difference that the treatment causes, averaged over the population. Identifying τ requires consistency, positivity $0 < \Pr(T = 1 \mid X) < 1$, and ignorability $\{Y(0), Y(1)\} \perp T \mid X$. Ignorability requires every confounder to appear in X. This rarely holds in practice.

2.2 Sensitivity analysis

When a confounder is unobserved, ignorability fails. Sensitivity analysis then asks how strong that confounder would have to be to overturn the estimated effect. The marginal sensitivity model (MSM) of Tan (2006) makes this precise on the propensity scale, and is widely used for inverse-propensity-weighting analyses. Write $e(x) = \Pr(T = 1 \mid X = x)$ for the observed propensity score and $e^\star(x, u) = \Pr(T = 1 \mid$

$X = x, U = u$) for the full-information one. The MSM bounds their odds ratio by a single parameter $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \;\leq\; \mathrm{OR}(x,u) \;\leq\; \Gamma, \tag{1}$$

where

$$\mathrm{OR}(x,u) \;=\; \frac{e(x)\,/\,(1-e(x))}{e^{\star}(x,u)\,/\,(1-e^{\star}(x,u))}$$

is the odds ratio, in the orientation of Jin et al. (2022). The bound holds for all (x,u). At $\Gamma = 1$ we recover ignorability. As Γ grows, the partially identified set of treatment effects widens.

2.3 Informal benchmarking

Choosing Γ is the main practical difficulty with the MSM. *Informal benchmarking* (Imbens, 2003) addresses this by gauging the unobserved confounder from each observed covariate in turn. For covariate $i \in \{1, \ldots, p_X\}$, we refit the propensity score without it, giving $\hat{e}^{(-i)}$. The odds ratio

$$\hat{r}_{i,j} \;=\; \frac{\hat{e}(X_j)\,/\,(1-\hat{e}(X_j))}{\hat{e}^{(-i)}(X_j)\,/\,(1-\hat{e}^{(-i)}(X_j))}, \tag{2}$$

at each observation j measures how much that covariate moves the treatment odds. We take the largest ratio as covariate i's contribution to Γ, and iterating over covariates yields a plausible interval $[\hat{\Gamma}_{\mathrm{low}}, \hat{\Gamma}_{\mathrm{high}}]$. The identifying assumption is that no unobserved confounder is stronger than the strongest observed one. The procedure inherits the pointwise structure of the MSM, so a single extreme observation can dominate $\hat{\Gamma}$, and it is sensitive to positivity, since the ratios in (2) diverge where the fitted propensity is already near zero or one.

2.4 The f-sensitivity model

The pointwise bound (1) is conservative when confounding is large in a small region but small on average. Jin et al. (2022) give the example of a Gaussian confounder acting through a linear logit. Its odds ratio is unbounded over the tails of U, so no finite Γ admits it, even though its effect on τ is small and finite. In this case the MSM gives wide bounds while the true bias in τ is small.

The f-sensitivity model replaces the pointwise constraint (1) with an average one. For a convex $f\colon \mathbb{R}_{>0} \to \mathbb{R}$ with $f(1) = 0$, the (f,ρ)-*selection condition* of Jin et al. (2022) requires, for P-almost every x,

$$\max\Big\{\, \mathbb{E}\big[\, f(\mathrm{OR}(X,U)) \,\big|\, X = x,\, T = 1\big]\,, \\ \mathbb{E}\big[\, f\big(\mathrm{OR}(X,U)^{-1}\big) \,\big|\, X = x,\, T = 0\big] \,\Big\} \;\leq\; \rho, \tag{3}$$

where $\rho \geq 0$ is the sensitivity parameter. Each side is an f-divergence that is zero under ignorability. The maximum keeps the bound symmetric across arms, following Jin et al. (2022). Up to identifiability conditions, each side equals a conditional f-divergence between the counterfactual and observed outcome distributions in that arm. The condition combines two quantifiers asymmetrically. Within each x it bounds an expectation over the confounder U, but it must hold for P-almost every x. This is the precise sense in which the model relaxes the MSM "on average". The averaging is over U within a covariate value, while the values themselves remain worst-case. It is also why the benchmark of Section 3.1 takes the *maximum* over covariates of a per-covariate average.

Different f give different divergences. This paper uses the Kullback–Leibler choice $f(t) = t \log t$, which penalises the log-ratio of densities, so it reduces the influence of extreme values more than alternatives such as χ^2 ($f(t) = (t-1)^2$). The Cressie–Read family interpolates between these standard cases (Jin et al., 2022). The f-sensitivity model is thus a distribution-aware generalisation of the MSM, with ρ interpolating between ignorability and no information through a divergence budget rather than a uniform ratio bound.

Two differences matter for benchmarking. First, Γ is a worst-case odds ratio over every (x,u), while ρ is a worst case *over* x of an average *over* U, so the two are not on a common scale. Second, this gap can be large. Heterogeneous confounding that is large on a small subset and near unity elsewhere can satisfy a modest ρ yet need an arbitrarily large Γ. The next section applies these differences and introduces an adaptation: it keeps the drop-one procedure and maximum across covariates, but summarises each covariate's arm shift by an f-divergence rather than its largest ratio.

3 Methodology

This section defines the procedure and the simulated scenarios used to test it. The differences of Section 2.4 mean we cannot reuse the MSM leave-one-out step unchanged, because it returns a worst-case odds ratio while ρ bounds a divergence. We keep the drop-one-covariate idea but, for each covariate, summarise the pooled treated/control shift by a symmetric-KL divergence rather than a worst-case ratio, taking the

strongest covariate as the benchmark (Section 3.1). We then build two families of confounding distributions to test it (Section 3.2) and a data-generating process with known ground truth (Section 3.4) that simulates a hidden confounder whose induced divergence is available in closed form, so we can check that the benchmark recovers it (Section 3.3). All quantities are defined in both the f-sensitivity (ρ) and the marginal-sensitivity (Γ) form, so the two frameworks are read on identical data.

Algorithm 1 The f-sensitivity benchmark $\hat{\rho}_{\text{bench}}$.

Require: analysis (X, T)
1: **for** each observed covariate X_j **do**
2: form pooled arm-conditional p_j^1, p_j^0 over the whole sample
3: $\hat{\rho}_j \leftarrow$ symmetric KL (4)
4: **end for**
5: $\hat{\rho}_{\text{bench}} \leftarrow \max_j \hat{\rho}_j$
6: **return** $\hat{\rho}_{\text{bench}}$

3.1 Setup

The benchmark $\hat{\rho}_{\text{bench}}$ is calculated only from the observed covariates and treatment. Whether it is accurate (whether it recovers the divergence a true hidden confounder would induce) is checked in simulation once the scenarios are in place (Section 3.3).

Benchmark. Let X_j be a discrete observed covariate. We drop it and compare the distribution of X_j across the treated and control arms over the whole sample, $p_j^1 = \Pr(X_j \mid T = 1)$ and $p_j^0 = \Pr(X_j \mid T = 0)$. We write $p_j^t(v) = \Pr(X_j = v \mid T = t)$ for the value at each v, so p_j^t collects these over all values of X_j. The divergence that dropping X_j induces between the arms is (3) evaluated at $U = X_j$ with no conditioning covariate. Each side is the expectation of f over the values of X_j under that arm's law p_j^t. With X_j in the confounder role, $\text{OR}(X_j)$ is the odds ratio (1) with the marginal $\Pr(T = 1)$ as the observed propensity and the per-value $\Pr(T = 1 \mid X_j)$ as the full-information one. Evaluating (3) at $U = X_j$ applies f to this odds ratio and averages it under each arm's law. The symmetric maximum is

$$\hat{\rho}_j \;=\; \max\Big\{ \mathbb{E}_{T=1}\big[\, f(\text{OR}(X_j))\big]\,, \;\; \mathbb{E}_{T=0}\big[\, f\big(\text{OR}(X_j)^{-1}\big)\big] \;\Big\}. \tag{4}$$

For the KL generator $f(t) = t \log t$, each expectation is a Kullback–Leibler divergence between the arm laws. Each $\hat{\rho}_j$ is the divergence, over the values of X_j, that dropping X_j induces across the sample. The overall benchmark is the strongest such covariate,

$$\hat{\rho}_{\text{bench}} \;=\; \max_j \hat{\rho}_j. \tag{5}$$

The benchmark keeps the MSM informal-benchmarking construction (Section 2.3) and only changes the scale it is read on. Equation (4) takes the same odds-ratio object through $f(t) = t \log t$, replacing the worst-case ratio over the values of X_j by a divergence while keeping the maximum over covariates. Taking the strongest observed covariate as the reference rests on the assumption that no unobserved confounder is stronger than it (Section 2.3), and matches the worst-case character of ρ, required to hold for every x rather than only on average (Section 2.4). The MSM benchmark $\hat{\Gamma}_{\text{bench}} = \max_j \hat{\Gamma}_j$, computed for comparison on the same drop-one procedure, instead records the worst-case odds ratio over the values of X_j. The trade-off is that p_j^t isolates X_j alone only when the covariates are uncorrelated, a condition the process of Section 3.4 guarantees. Propensities entering $\hat{\Gamma}$ are clipped at $\varepsilon = 10^{-3}$, and the benchmark is averaged over seeds.

3.1.1 Algorithm

Algorithm 1 states the procedure. From the observed covariates and treatment indicator (X, T), it forms each covariate's treated- and control-arm distributions p_j^1, p_j^0, computes their symmetric KL divergence $\hat{\rho}_j$ (Eq. 4), and returns the largest as $\hat{\rho}_{\text{bench}}$. It makes a single pass over the covariates.

3.2 Scenarios

The premise of the f-sensitivity model is that a worst-case ratio and a divergence can separate. We implement the contrasting $\text{OR}(x, u)$ curves of Jin et al. (2022) on data from our process (Figure 1), using two scenario families. Each family holds one framework's reading fixed while moving the other. In the first, a rare narrow "spike" and a common "broad" bump are tuned to the same worst-case ratio Γ but differ several-fold in divergence ρ, so f-sensitivity distinguishes distributions the MSM cannot. In the second, a smooth body and the same body plus a rare deep tail-spike are tuned to a similar ρ but differ sharply in Γ, so the MSM responds to a concentrated group that adds little to the divergence ρ. Together these let us check that $\hat{\rho}_{\text{bench}}$ tracks the divergence. It distinguishes distributions the MSM reads as identical, and it stays flat where the MSM responds to a rare group. This is how an f-sensitivity budget should

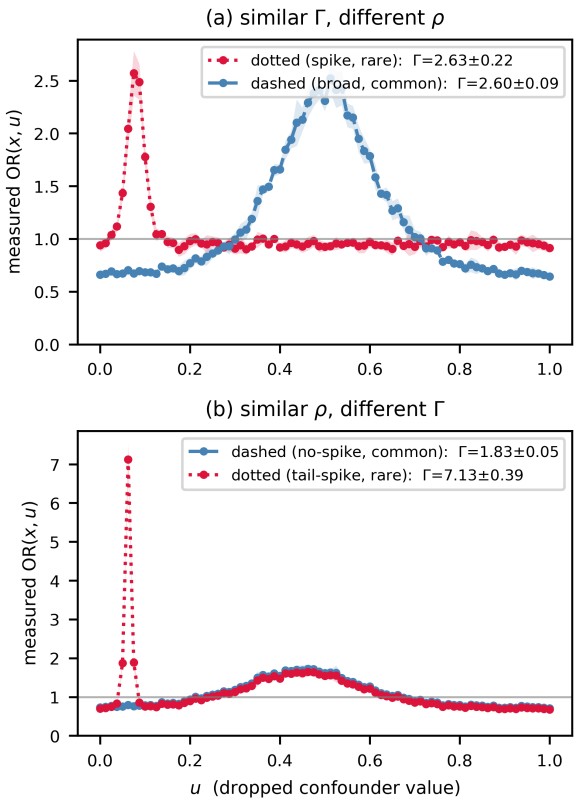

Figure 1: Empirical $\mathrm{OR}(x, u)$ curves measured from data generated by two multi-valued confounders. *(a)* Γ a rare narrow "spike" and a common "broad" bump reach almost the same worst-case ratio Γ, but have very different divergence. *(b)* ρ a common body and the same body with an added rare deep tail-spike have a similar divergence, but Γ is much larger on the tail-spike. The MSM does not distinguish the top pair and is sensitive to the bottom one, while f-sensitivity shows the reverse.

behave, and the experiments of Section 4 confirm it.

3.3 Recovering the true divergence from a known confounder

The benchmark reads the observed covariates to gauge the unobserved one. What decides its accuracy is whether it recovers the divergence a true hidden confounder induces. In the process of Section 3.4 the confounder U is known and the propensity $\Pr(T = 1 \mid X, U)$ is a fixed logistic, so the divergence U induces is available in closed form, the per-covariate divergence (4) computed with the true confounder U in place of the observed X_j.

The benchmark recovers the truth when $\hat{\rho}_{\text{bench}} \geq \rho_{\text{true}}$, making the informal-benchmarking assumption, that no unobserved confounder is stronger than the strongest observed one (Section 2.3), checkable. The same comparison runs on the MSM scale with $\hat{\Gamma}_{\text{bench}}$ and Γ_{true}, and Section 4.3 reports both.

3.4 Data generation process

The experiments use two related data-generating processes. The benchmark-only experiments (Sections 4.1–4.2) need covariates and a treatment alone. The recovery experiment (Section 4.3) adds the unobserved confounder U and checks the benchmark against the divergence U induces. Both draw n observations per replication in causal order and estimate propensities non-parametrically, so the readings see the empirical distribution rather than a fitted model.

3.4.1 Benchmark-only experiments with multi-valued covariates

These experiments build each multi-valued covariate on a grid of $[0, 1]$ and draw treatment from a per-value propensity table, $T \sim \text{Bernoulli}\big(e(X_j)\big)$, with $e(\cdot)$ shaped to the scenario (spike, broad bump, plain body, or body-plus-tail-spike). A deeper local dip in e raises the realised odds ratio at that value. The worst-case-versus-average comparison (Section 4.1) uses a single covariate on a grid of 81 values, fine enough to resolve the narrow features it relies on, with a uniform distribution over them. The multi-covariate spike and heterogeneous sweeps (Section 4.2) use three independent such covariates on a coarser grid of 25 with a peaked distribution. Because the benchmark reads only (X, T), these experiments use no hidden confounder U. There is no confounding to recover, only the benchmark's behaviour to read.

3.4.2 Recovery experiment with covariates and a hidden U

The recovery experiment adds an unobserved confounder U, drawn in causal order (U and X, then T). Dropping U before estimation leaves only the observed covariates for the benchmark to read, while the known U supplies the ground-truth divergence to recover. The exogenous variables are independent, with $U \sim \text{Bernoulli}(u)$ and binary $X_j \sim \text{Bernoulli}(p_j)$. So adjusting for X removes none of U's influence, and what the benchmark reads reflects only the confounding the procedure is meant to bound.

Treatment follows a logistic model

$$\Pr(T = 1 \mid X, U) = \sigma\Big(t_0 + \textstyle\sum_j \beta_j X_j + t_u\, U\Big), \quad (6)$$

so each covariate confounds treatment through β_j and U through t_u. The coefficients are kept modest

so that every occupied group keeps both treatment arms. Because both the benchmark and the true divergence ρ_{true} (Section 3.3) read only the treatment assignment, the process needs no outcome.

4 Experimental Setup and Results

We test the procedure on simulated data with known ground truth, varying the data-generating process and reading the benchmark as the main quantity. The three experiments take the sub-questions of the introduction in turn: the first (Section 4.1) asks whether the worst-case and average-divergence scales separate at all (sub-question ii), the second (Section 4.2) whether that separation survives a genuine maximum over covariates and where the benchmark fails (sub-questions iii and v), and the third (Section 4.3) whether the benchmark recovers the divergence a known confounder induces (sub-question iv). Together they test whether $\hat{\rho}_{\text{bench}}$ behaves like an average-divergence budget and reads the confounding faithfully. The first runs on a single covariate, isolating the per-covariate statistic $\hat{\rho}_j$ (the building block the benchmark takes the maximum over) apart from the drop-one aggregation, which is vacuous with one covariate; the last two use several covariates, where the benchmark is a genuine maximum $\hat{\rho}_{\text{bench}} = \max_j \hat{\rho}_j$ and the covariate that attains it can change. Each subsection states its question, its setup, what we expect and why, and what we find.

4.1 The worst-case versus average gap

This experiment addresses sub-question (ii). On the same data, do the MSM's worst case over (x, u) and f-sensitivity's worst case over x of an average over U separate, with each ranking covariates that the other treats as alike. We build the two scenario families of Section 3.2 to force the answer (Figure 1). One pair is a rare narrow spike and a common broad bump, tuned to the same worst-case ratio. The other pair is a plain body and the same body plus a deep tail-spike, tuned to a similar average divergence. Suppose $\hat{\rho}_{\text{bench}}$ reads the average while $\hat{\Gamma}$ reads the worst case. Then the first pair should match in Γ but differ several-fold in ρ, and the second should match in ρ but differ several-fold in Γ. The spike and broad bump reach almost the same worst-case ratio ($\hat{\Gamma}_j \approx 2.6$ for both) but differ about five-fold in average divergence. The plain and tail-spike bodies differ only about 1.5-fold in average divergence but almost four-fold in worst-case ratio ($\hat{\Gamma}_j$ of 1.8 and 7.1). So $\hat{\rho}_{\text{bench}}$ tracks the average divergence. It separates distributions that the worst-case Γ treats as identical, which is the behaviour we want from an f-sensitivity budget.

4.2 Benchmarking across several covariates

The single-covariate experiment of Section 4.1 characterises $\hat{\rho}_j$ on fixed scenarios; this one sweeps a covariate continuously and asks whether those behaviours survive when the benchmark is a genuine maximum over several covariates, $\hat{\rho}_{\text{bench}} = \max_j \hat{\rho}_j$, whose attaining covariate can change (sub-question iii). It also exposes where the benchmark breaks down (sub-question v). Using three independent multi-valued covariates X_0, X_1, X_2, we run two sweeps. In the first, all three bumps share a peak height, X_1, X_2 are fixed, and X_0's bump moves from the sparse tail of its distribution into the denser part. Because the divergence weights each covariate value by its probability, this is the regime in which it under-reports confounding concentrated on low-probability values, the case sub-question (v) targets. In the second, X_1, X_2 carry only a body bump while X_0 carries the same body plus a tight spike in its sparse tail (mass 0.011) whose height grows; there the worst case climbs while the probability-weighted average barely moves.

We expect each benchmark to move only when its own scale registers the change, with the binding covariate switching to X_0 once it overtakes the anchors. The matched-peak bump leaves the worst case unchanged, so $\hat{\Gamma}_{\text{bench}}$ should hold while $\hat{\rho}_{\text{bench}}$ rises with the mass X_0 covers; the growing spike should do the reverse. Both hold (Figures 2 and 3). In the first sweep $\hat{\Gamma}_{\text{bench}}$ stays near 5.7 while $\hat{\rho}_{\text{bench}}$ rises from 0.036 to 0.186, the binding covariate switching from an anchor to X_0; in the second $\hat{\Gamma}_{\text{bench}}$ climbs from 2.3 to 6.0 while $\hat{\rho}_{\text{bench}}$ holds near 0.047. There a large but rare spike sets the worst case while leaving the probability-weighted average almost unchanged, matching the right panel of Figure 1 in Jin et al. (2022), where the separation appears across the shapes tested and not only the fixed pairs of Section 4.1. The weak point is the low-overlap end of the first sweep, where with X_0's bump in the sparse tail, $\hat{\rho}_{\text{bench}}$ reads only 0.036 despite the matched worst case $\hat{\Gamma}_{\text{bench}} \approx 5.7$, under-reporting the strongly but locally confounded X_0 until its bump gains mass. The fixed-scenario contrasts of Section 4.1 carry into the full benchmark, and the maximum tracks the binding covariate on each scale.

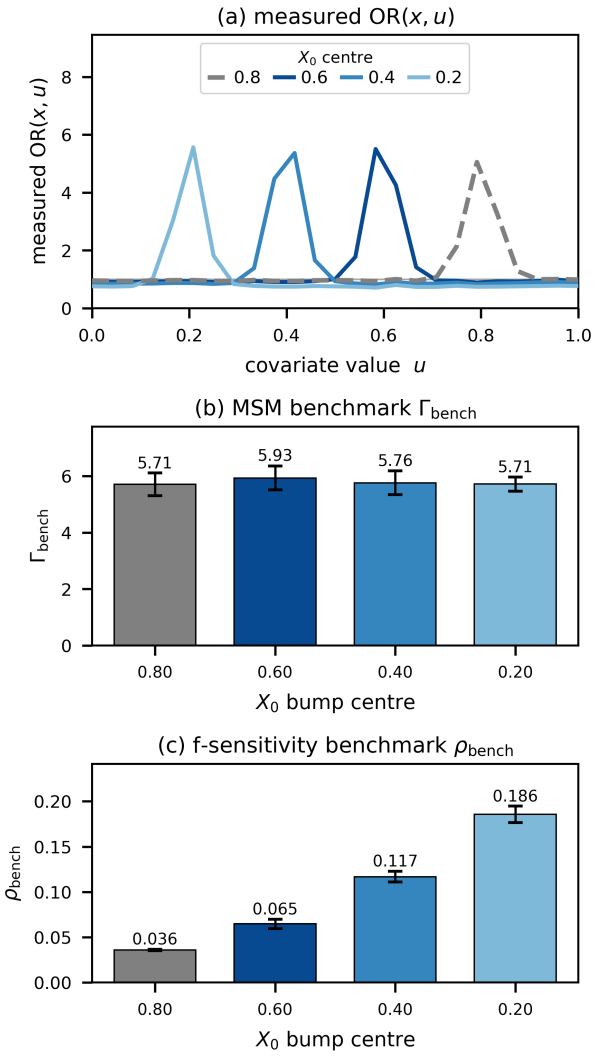

Figure 2: The f-sensitivity benchmark rises while the MSM benchmark stays matched, across three covariates. X_1 and X_2 are anchored at covariate value 0.8; X_0's propensity bump, of fixed peak height, marches left into the denser part of its distribution (overlap $0.03 \to 0.16$). Mean $\pm$ std, five seeds. *(a)* measured $\mathrm{OR}(x, u)$, X_0's bump slides left; its 0.8 position coincides with the X_1, X_2 anchor (dashed grey). *(b)* $\hat{\Gamma}_{\mathrm{bench}} = \max_j \hat{\Gamma}_j$ stays near 5.7, since all bumps share a peak height. *(c)* $\hat{\rho}_{\mathrm{bench}} = \max_j \hat{\rho}_j$ rises from 0.036 to 0.186 as X_0 covers more mass; the benchmark's binding covariate switches from an anchor to X_0 at centre 0.6.

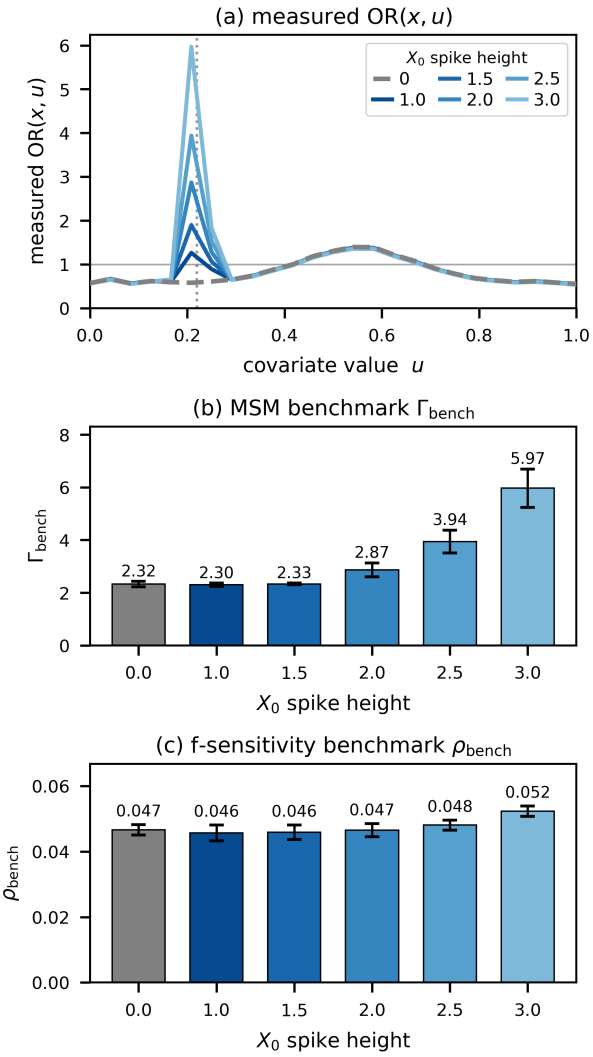

Figure 3: The mirror case, where the MSM benchmark climbs while the f-sensitivity benchmark stays flat. X_1 and X_2 carry a shared body bump; X_0 carries the same body plus a tight spike in the sparse tail (mass 0.011) whose height grows. Mean $\pm$ std, five seeds. *(a)* measured $\mathrm{OR}(x, u)$, the shared body and X_0's growing spike. *(b)* $\hat{\Gamma}_{\mathrm{bench}}$ climbs from 2.3 to 6.0 once X_0's spike clears the body's worst-case ratio; the binding covariate switches to X_0 at spike height 2.0. *(c)* $\hat{\rho}_{\mathrm{bench}}$ stays near 0.047 (rising only to 0.052), since the spike carries little mass.

4.3 Recovering the true divergence

This experiment addresses sub-question (iv). Does the benchmark produce an accurate estimate, recovering the divergence that the hidden confounder actually induces? Because U is simulated and the propensity is known, the true value ρ_{true} it induces is available in closed form (Section 3.3). This gives a ground truth to read the benchmark against. We use four scenarios. Each fixes the shapes of three binary covariates and a hidden U. For each scenario we compute the benchmark over the observed covariates ($\hat{\rho}_{\mathrm{bench}} = \max_j \hat{\rho}_j$, seed-averaged) and the

analytic truth ρ_{true}, with the MSM analogues $\hat{\Gamma}_{\text{bench}}$ and Γ_{true} on the same data. The scenarios span weak covariates with weak U, a single rare-strong covariate among weak ones, a uniformly common-moderate set, and strong covariates with strong U.

Suppose the informal-benchmarking assumption holds, meaning no unobserved confounder is stronger than the strongest observed one. Then the benchmark should cover the truth, $\hat{\rho}_{\text{bench}} \geq \rho_{\text{true}}$, and likewise on the MSM scale (Figure 4). In all four scenarios the benchmark read from the observed covariates upper-bounds the divergence that U actually induces, on both the ρ and the Γ scale. The cover is tightest when confounding is homogeneous across the covariates (strong covariates, strong U, $\hat{\rho}_{\text{bench}} = 0.20$ against $\rho_{\text{true}} = 0.16$). It is loosest when one observed covariate is far stronger than U (one rare-strong covariate, 0.22 against 0.06), where the strongest observed covariate sets a benchmark well above the hidden confounder's divergence. So the benchmark recovers the true ρ across the scenarios, in the sense of conservatively covering it. This is the accuracy that sub-question (iv) asks for.

5 Responsible Research

Reproducibility. All evidence comes from simulation, which avoids the usual data-privacy and consent concerns but raises ones about reproducibility. Every data-generating process is fixed by a random seed, and each reported quantity is averaged over a contiguous range of seeds, so a single integer determines the data. Generation runs on one worker to keep the draws reproducible. We fit propensities non-parametrically with one clipping constant $\varepsilon = 10^{-3}$, and a group enters the benchmark only above explicit minimum arm and cell counts. We release these thresholds, the seed counts, and the pinned Python stack with the code at `https://github.com/ASlics/bsc-informal-benchmarking-f-sensitivity-model`. Reproducibility here means matching the reported distribution of estimates across seeds, not a single-seed point estimate, which would overstate precision.

Integrity. A simulation study can be tuned to favour a procedure by choosing data on which it does well. To guard against this, we evaluate the benchmark both where it behaves as intended (Sections 4.1 and 4.2) and where it fails (Section 4.2, confounding in a low-density region), and report the recovery check on both the f-sensitivity and the MSM scale, not as an advantage for either. The comparison is framed narrowly, against the ignorability analysis at $\rho = 0$ and the true divergence the simulated confounder induces, not as a general claim of superiority. Because the confounder is synthetic and known, the conclusions describe the studied settings rather than endorsing the benchmark for real decisions, where a benchmark that under-estimates confounding would fail toward false reassurance.

Use of AI assistance. We used Anthropic's Claude (via the Claude Code interface) for code drafting and debugging, LaTeX and language editing, and manuscript copy-editing. It was not used to generate research ideas, derivations, experimental designs, or results: all modelling decisions, derivations, experimental designs, and interpretations are the author's own. We reviewed all AI-assisted text and code before including it. No part of the empirical evidence was produced by autonomous AI experimentation without author review.

6 Discussion

The experiments support one main claim, that informal benchmarking transfers to the f-sensitivity model, producing from the observed covariates alone a stable benchmark $\hat{\rho}_{\text{bench}}$ that behaves as an average-divergence budget should. This section develops that claim and then states the limits of the evidence.

The two models read confounding differently. The two benchmarks summarise a covariate differently, one by its average shift over the covariate's values, the other by its worst value, so they can rank the same covariate in opposite orders (the spike sweep of Section 4.2). Neither reading is incorrect on its own terms. They answer different questions. The MSM asks whether any single group could overturn the effect, while f-sensitivity asks whether the average confounding could. Which is appropriate is a substantive choice. If a plausible confounder is strong on a small subgroup and absent elsewhere, the f-sensitivity reading is the smaller. If it behaves like a common moderate covariate, the two move together. The recovery check confirms this. The gap between $\hat{\rho}_{\text{bench}}$ and ρ_{true} is widest exactly when one covariate is much stronger than the rest.

Limitations. Three limitations qualify these results. First, the benchmark takes the maximum over covariates, as informal benchmarking does (no unobserved confounder stronger than the strongest

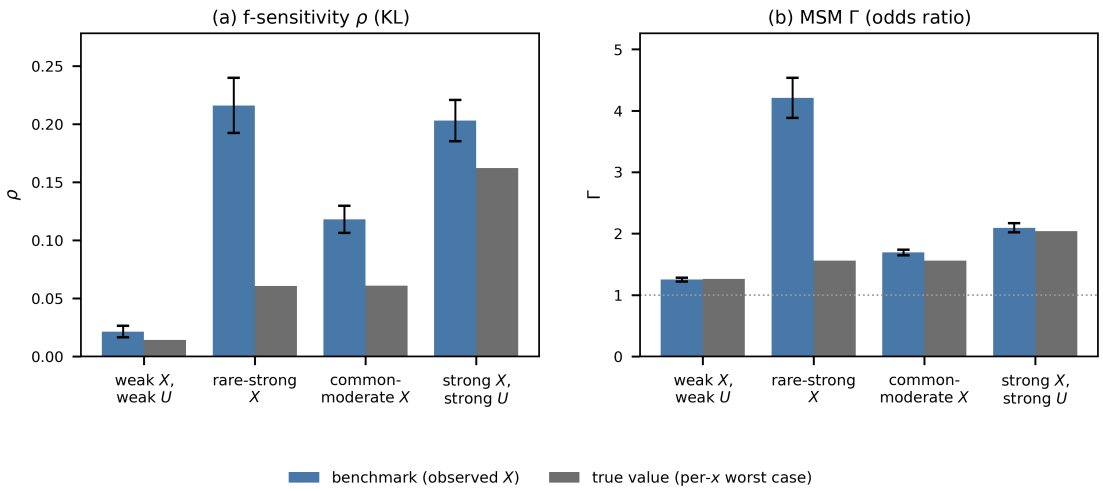

Figure 4: The benchmark recovers the true divergence that the hidden confounder induces, across four covariate scenarios. *(a)* f-sensitivity ρ (KL), the benchmark over the observed covariates $\hat{\rho}_{\text{bench}}$ (mean $\pm$ std over 15 seeds) against the analytic truth ρ_{true}, taken as the per-x worst case that the selection condition demands. *(b)* the MSM analogue on the Γ scale. In every scenario $\hat{\rho}_{\text{bench}} \geq \rho_{\text{true}}$ and $\hat{\Gamma}_{\text{bench}} \geq \Gamma_{\text{true}}$. The benchmark covers the hidden confounding, conservatively, with the cover tightest under homogeneous confounding and loosest when one observed covariate is much stronger than U.

observed one), and mirroring $\hat{\Gamma}_{\text{bench}}$. It is therefore driven by a single covariate. In finite samples its maximising covariate can change from seed to seed. The error bars in Figure 4 show the resulting spread stays modest. Second, the benchmark is pooled, reading each covariate's shift over the whole sample as if that covariate were uncorrelated with the others, whereas the (f, ρ) condition is a per-x worst case. The two agree when the covariates are uncorrelated, as the process ensures, but can separate when confounding is strongly heterogeneous across covariate values. Third, the experiments use binary covariates, a non-parametric propensity estimator, and linear-on-logit confounding. Under continuous covariates and fitted propensities, the benchmark would inherit the estimator's bias. Because the confounder is synthetic and known, the results characterise the studied settings rather than endorsing the benchmark for a real application.

7 Conclusions and Future Work

Informal benchmarking transfers to the f-sensitivity model through a benchmark $\hat{\rho}_{\text{bench}}$, the symmetric KL that the strongest dropped covariate induces between the treatment arms, read from the observed covariates alone with no outcome model or solver. It is stable across seeds and behaves as an average-divergence budget should, staying flat under a rare spike while the MSM's worst-case reading climbs. On simulated data with a known hidden confounder it recovers the induced divergence, covering the true ρ (and the true Γ on the MSM scale) in all four scenarios. The cover is conservative, tightest under homogeneous confounding and loosest when one observed covariate is much stronger than the hidden one, so the informal-benchmarking assumption (no unobserved confounder stronger than the strongest observed one) holds throughout. Its one weakness is that the mass-weighted divergence under-reports confounding concentrated in a low-density region of a covariate's range, the case the multi-covariate experiment isolates (Section 4.2).

Future work. The pooled benchmark reads each covariate's shift over the whole sample, treating the covariates as uncorrelated. A covariate-adjusted aggregation that avoids the small-cell bias of naive stratification would extend it to correlated covariates and might admit a consistency guarantee that the present summary lacks. Other directions include continuous covariates with fitted, regularised propensities, generators other than KL across the Cressie–Read family, and extending the recovery check to settings where the true divergence must itself be estimated.

References

Baitairian, J.-B., Sebastien, B., Jreich, R., Katsahian, S., and Guilloux, A. (2025). Calibrating confounding strength in sensitivity models for weighting estimators: A comparative review and a new method.

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., and Syrgkanis, V. (2024). Long story short: Omitted variable bias in causal machine learning.

Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B*, 82(1):39–67.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132.

Jin, Y., Ren, Z., and Zhou, Z. (2022). Sensitivity analysis under the f-sensitivity models: A distributional robustness perspective.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, 15(5):1–46.