Automatic debiased machine learning and sensitivity analysis for sample selection models

Jakob Bjelac JAKOB.BJELAC@OUTLOOK.COM
Putlitzstraße 2, 10551 Berlin, c/o Valdez

Victor Chernozhukov VCHERN@MIT.EDU
Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142, USA

Phil-Adrian Klotz KLOTZ@DICE.HHU.DE
Düsseldorf Institute for Competition Economics, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, North Rhine–Westphalia, Germany

Jannis Kueck KUECK@DICE.HHU.DE
Düsseldorf Institute for Competition Economics, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, North Rhine–Westphalia, Germany

Theresa M. A. Schmitz THERESA.SCHMITZ@HHU.DE
Chair of Statistics and Econometrics, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, North Rhine–Westphalia, Germany

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

In this paper, we extend the Riesz representation framework to causal inference under sample selection, where both treatment assignment and outcome observability are non-random. Formulating the problem in terms of a Riesz representer enables stable estimation and a transparent decomposition of omitted variable bias into three interpretable components: a data-identified scale factor, outcome confounding strength, and selection confounding strength. For estimation, we employ the ForestRiesz estimator, which accounts for selective outcome observability while avoiding the instability associated with direct propensity score inversion. We assess finite-sample performance through a simulation study and show that conventional double machine learning approaches can be highly sensitive to tuning parameters due to their reliance on inverse probability weighting, whereas the ForestRiesz estimator delivers more stable performance by leveraging automatic debiased machine learning. In an empirical application to the gender wage gap in the U. S., we find that our ForestRiesz approach yields larger treatment effect estimates than a standard double machine learning approach, suggesting that ignoring sample selection leads to an underestimation of the gender wage gap. Sensitivity analysis indicates that implausibly strong unobserved confounding would be required to overturn our results. Overall, our approach provides a unified, robust, and computationally attractive framework for causal inference under sample selection.

Keywords: Sample Selection, Automatic Debiased Machine Learning, Riesz Representation, ForestRiesz, Sensitivity Analysis

© 2026 J. Bjelac, V. Chernozhukov, P.-A. Klotz, J. Kueck & T.M.A. Schmitz.

1. Introduction

In many empirical studies, researchers face the challenge that outcomes are only observed for a subset of the sample population. Returns to education studies observe wages only for employed individuals. Job training evaluations miss earnings data for unemployed participants. Clinical trials lose patients before outcome measurement and also educational interventions suffer when students do not take standardized tests. This phenomenon, commonly referred to as sample selection or outcome attrition, complicates the estimation of causal effects (Heckman, 1976, 1979; Hausman and Wise, 1979; Little, 1995). The problem becomes even more complex when treatment assignment is itself non-random. In such cases, researchers confront what Bia et al. (2024) describe as the "double selection problem", involving both selection into treatment and selection into outcome observability. Standard methods for confounding adjustment, such as regression or propensity score weighting, fail when outcomes are selectively missing. Even inverse probability weighting, which addresses treatment selection, requires modification to handle missing outcomes (Robins et al., 1994; Hernán et al., 2004). The machine learning literature offers powerful tools for high-dimensional covariate adjustment but introduces new challenges. In particular, regularization inherent in machine learning estimators can induce bias that invalidates standard inference procedures unless appropriate orthogonality conditions are imposed (Chernozhukov et al., 2018). Bia et al. (2024) address this issue by deriving a Neyman-orthogonal score function for treatment effect estimation in the presence of sample selection. Dolgikh and Potanin (2025) also propose double machine learning estimators for treatment effect estimation in the multivariate sample selection model with ordinal selection equations. An alternative approach is provided by the Riesz representation theorem. Instead of relying on Neyman-orthogonal score functions, target parameters can be characterized through unique weighting functions called Riesz representers (Chernozhukov et al., 2022d). The Riesz framework offers several advantages: it avoids unstable propensity score inversions, enables direct estimation via variational or adversarial methods, and naturally accommodates sensitivity analysis. Existing applications include the estimation of the Average Treatment Effect (ATE) and other policy-relevant causal parameters in settings without selection (Chernozhukov et al., 2022d) and the estimation of the Average Treatment Effect on the Treated (ATT) in Difference-in-Differences models (Bach et al., 2025).

This paper extends the Riesz representation methods to sample selection models. We show that the ATE identified by Bia et al. (2024) via efficient scores also admits identification through a Riesz representation. The corresponding representer takes the form of inverse probability weights that adjust simultaneously for treatment assignment and sample selection. The Riesz representer framework is particularly useful for analyzing bias induced by unobserved selection confounding. It yields an interpretable decomposition of the omitted variable bias. Building on Cinelli and Hazlett (2020) and Chernozhukov et al. (2022a), we express the bias as the product of three terms: (i) a scale factor identifiable from observed data, (ii) the strength of confounding in the outcome equation, and (iii) the strength of confounding in the selection equation. This decomposition delivers sharp bounds on the magnitude of bias without requiring the specification of the full joint distribution of unobservables. A key insight is that observed covariates provide natural benchmarks for calibrating these sensitivity parameters (Imbens, 2003; Altonji et al., 2005; Oster, 2019). In our simulation study, we investigate the finite-sample behavior of the proposed ForestRiesz estimator and find that it performs well in finite samples when estimating the ATE. As an empirical contribution, we study the gender wage gap in the U. S. using data from the American Community Survey. We find that

our ForestRiesz approach yields larger treatment effect estimates than a standard double machine learning approach which does not account for sample selection. This suggests that ignoring sample selection leads to an underestimation of the gender wage gap, as wage reporting behavior differs systematically between female and male respondents.

2. Identification under Confounding and Sample Selection

Estimation of treatment effects is fundamental to empirical research in economics, medicine, and the social sciences. This section introduces the Average Treatment Effect (ATE) within the potential outcomes framework and examines the problem of sample selection, which occurs when outcome data are missing for some units in the analysis.

2.1. Defining Causal Effects: The Potential Outcomes Framework

To formally define causal effects, we rely on the Potential Outcomes framework, often associated with Rubin (1974, 1977). Let D be a variable that represents the treatment status assigned to an individual unit i. For clarity, we consider a binary treatment where $D_i = 1$ if unit i receives the treatment and $D_i = 0$ if unit i receives the control, though the framework readily extends to multiple discrete treatments $d \in \{0, 1, ..., Q\}$. For each unit i, we define two potential outcomes: $Y_i(1)$ is the outcome that unit i would have experienced under treatment ($D_i = 1$), while $Y_i(0)$ is the outcome it would have experienced under control ($D_i = 0$). We assume SUTVA: unit i's potential outcomes are unaffected by other units' treatment assignments, and $Y_i(1)$ and $Y_i(0)$ are well-defined for each unit (Rubin, 1980). Given the impossibility of observing individual treatment effects directly for a single unit, empirical research typically focuses on estimating average causal effects across a population or subpopulation. The most common target parameter is the Average Treatment Effect (ATE) for the entire population:

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)],$$

where the expectation $\mathbb{E}[\cdot]$ is taken over the distribution of units in the population of interest.

2.2. The Sample Selection Problem

In many practical applications, the outcome variable Y is not observed for all units in the sample. This issue is known as sample selection, outcome attrition, or nonresponse. Let S be a binary indicator variable such that $S_i = 1$ if the outcome Y_i is observed for unit i, and $S_i = 0$ otherwise. If the mechanism determining whether the outcome is observed ($S = 1$) is related to the potential outcomes $Y(d)$ themselves, even after conditioning on treatment status D and covariates X, then the subsample for whom we observe the outcome is no longer representative of the full population concerning the outcome process. Simply performing an analysis on the selected sample without accounting for the selection mechanism can introduce sample selection bias. When non-random treatment assignment (violating $Y(d) \perp D$) occurs simultaneously with non-random sample selection (violating $Y(d) \perp S$), researchers face a so-called double selection problem, as emphasized by Bia et al. (2024). In this situation, valid estimation requires assumptions addressing both sources of potential bias. The first assumption invokes conditional independence of the treatment:

Assumption 1 *(Conditional Independence of the Treatment):*

$$Y(d) \perp D \mid X = x \text{ for all } d \in \{0, 1\} \text{ and } x \text{ in the support of } X.$$

By Assumption 1, no unobservables jointly affect the treatment and the potential outcomes conditional on covariates X. Analogous to how Assumption 1 addresses confounding in treatment assignment, specific assumptions are required to handle sample selection.

A common starting point is another selection-on-observables assumption, but applied to the selection process S. This is often termed the Missing-At-Random (MAR) assumption (Rubin, 1976) or conditional independence of selection. In the context of treatment evaluation, it means that conditional on observed variables (importantly, treatment status D and covariates X), the selection indicator S is independent of the potential outcomes:

$$Y(d) \perp S \mid D = d, X = x \quad \text{for all } d \in \{0, 1\} \text{ and } x \text{ in the support of } X. \tag{1}$$

This conditional independence of selection assumption states that, within groups defined by a specific treatment status d and covariate values x, whether an outcome $Y(d)$ is observed ($S = 1$) or missing ($S = 0$) does not depend on the potential outcome's value itself. Selection is allowed to depend on treatment D and covariates X, but not on any unobserved factors related to $Y(d)$ once D and X are accounted for. However, this might be violated in many real-world scenarios. Selection could depend on unobserved factors (denoted A) that also influence the potential outcome, even after conditioning on D and X. This is known as non-ignorable nonresponse or selection based on unobservables. In the following, we consider a much weaker condition, i. e., selection independence only holds if we could condition on the additional unobserved factors A. This leads to the following assumption:

Assumption 2 *(Conditional Independence with Unobservables):*

$$Y(d) \perp S \mid D = d, X = x, A = a \quad \textit{for all } d \in \{0, 1\} \textit{ and } x, a \textit{ in the support of } X \textit{ and } A.$$

Under Assumption 2, selection is independent of potential outcomes once we account for treatment status, observed covariates, and the unobserved selection confounding factors A. While we cannot observe A directly, we can apply the framework of Chernozhukov et al. (2022a) to provide sharp bounds on the size of the omitted variable bias that results from not observing A. When introducing unobserved confounders A in the selection process S, we need to specify how this affects the treatment assignment as well. We impose the following assumption:

Assumption 3 *(No Unobserved Confounding in the Treatment Assignment):*

$$A \perp D \mid X = x \quad \textit{for all } x \textit{ in the support of } X.$$

Assumption 3 restricts unobservables to affect outcome observability but not treatment assignment beyond X. This setting covers, e. g., stratified randomized controlled trials with attrition or nonresponse and observational studies in which rich covariates plausibly address treatment selection while outcome observability may remain selectively missing due to latent factors that also determine Y. The assumption may fail even after conditioning on X when unobservables jointly affect both treatment take-up and outcome observability, e. g., in program evaluations due to participants' unobserved motivation. Allowing latent factors to jointly affect D and S would require additional sensitivity components for treatment confounding; we leave this important extension to future work.

We also make the following assumption. Let $p_d(X) := \mathbb{P}(D = d \mid X)$ for $d \in \{0, 1\}$ and let $\pi_0(d, X, A) := \mathbb{P}(S = 1 \mid D = d, X, A)$.

Assumption 4 *(Common Support and Weak Overlap): Assume (i)* $p_d(X) > 0$ *and* $\pi_0(d, X, A) > 0$ *almost surely for* $d \in \{0, 1\}$*, and (ii) the inverse-propensity moments satisfy*

$$\mathbb{E}\left[\frac{1}{p_1(X)\pi_0(1, X, A)} + \frac{1}{p_0(X)\pi_0(0, X, A)}\right] < \infty. \tag{2}$$

We refer to Equation (2) as a *weak overlap* condition since it requires only integrability of inverse propensities (rather than uniform lower bounds). The first part of the assumption is a conventional common support condition, which ensures that treatment assignment is non-degenerate and the probability of selection is always non-zero for each conditioning value.

Weak Overlap. The Riesz approach requires square-integrability of the representer, $E[\alpha^2] < \infty$. In our setting, the representer includes the selection indicator S, so that its second moment collapses to an *average* inverse-probability scale. Concretely, for $d \in \{0, 1\}$,

$$E\left[\left(\frac{1\{D = d\}S}{p_d(X)\pi_0(d, X, A)}\right)^2 \middle| X, A\right] = \frac{E[1\{D = d\}S \mid X, A]}{p_d(X)^2\pi_0(d, X, A)^2} = \frac{1}{p_d(X)\pi_0(d, X, A)}.$$

Thus, Assumption 4 (ii) is sufficient; see also Appendix B. Additionally, we note that this aligns with the weak overlap condition implied by semi-parametric efficiency theory, cf. Newey (1994).

Further, we denote the conditional mean outcome by $\mu_d(X) = \mathbb{E}[Y|D = d, S = 1, X]$. Under Assumption 1, Assumption 4, and conditional independence of selection in Equation (1), the ATE is identified by:

$$\theta_0 = \mathbb{E}[\phi_1 - \phi_0]$$

with

$$\phi_d = \frac{\mathbf{1}\{D = d\} \cdot S \cdot [Y - \mu_d(X)]}{p_d(X) \cdot \pi_s(d, X)} + \mu_d(X) \tag{3}$$

being the efficient score function derived by Bia et al. (2024). Hence, the ATE is identified using outcomes Y from the selected sample ($S = 1$) and selection indicators S for all units. Intuitively, identification involves modeling the conditional outcome mean within the selected sample, $\mathbb{E}[Y|D = d, S = 1, X]$, and then appropriately adjusting or re-weighting based on estimates of the treatment propensity score $p_d(X) = \mathbb{P}(D = d|X)$ and the selection propensity score $\pi_s(d, X) = \mathbb{P}(S = 1|D = d, X)$.

3. Riesz Representers and Automatic Debiased Machine Learning

3.1. Neyman-Orthogonal Scores and the Role of the Riesz Representer

Many empirical problems now involve rich covariates. Machine-learning methods like Lasso, random forests, and neural networks can estimate nuisance functions such as conditional means and propensities in these settings. They achieve good prediction through regularization and model selection. These devices, however, typically introduce bias. If we plug a regularized estimate $\hat{g}$ into a target functional, the resulting estimator can inherit non-negligible bias and invalidate $\sqrt{n}$-consistent inference.

Debiased machine learning (DML) addresses this problem by using *Neyman-orthogonal* scores (Levit, 1975; Ibragimov and Has'minskii, 1979; Chernozhukov et al., 2018). In this framework, a score $\psi(W, \theta, g)$ identifies θ_0 via the moment condition

$$\mathbb{E}[\psi(W, \theta_0, g_0)] = 0$$

and is constructed so that small errors in g have only a second-order effect on the moment, where W denotes the data. Bia et al. (2024) derive a Neyman-orthogonal score for high-dimensional sample selection models (see Equation (3)) and use cross-fitting to obtain valid inference.

A complementary approach uses the Riesz representer. For many parameters of interest (including the ATE), we can write

$$\theta_0 = \mathbb{E}[m(W, g_0)],$$

where the map $g \mapsto \mathbb{E}[m(W, g)]$ is linear and continuous on a suitable function class. The Riesz Representation Theorem then yields a unique function α_0, called the Riesz representer, such that

$$\mathbb{E}[m(W, g)] = \mathbb{E}[\alpha_0(Z)\, g(Z)]$$

for all admissible functions g, where Z collects the arguments of g and α_0.

The condition $\mathbb{E}[\alpha_0(Z)^2] < \infty$ is closely linked to θ_0 having a finite semiparametric efficiency bound (Newey, 1994; Hirshberg and Wager, 2021; Chernozhukov et al., 2022c). In our setting, the efficient score of Bia et al. (2024) admits an analogous Riesz representation that combines treatment and selection propensity weights. The representer also leads to a generic orthogonal score. For target parameters of the form $\theta_0 = \mathbb{E}[m(W, g_0)]$ with $g_0(Z) = \mathbb{E}[Y \mid Z]$, consider

$$\psi(W, \theta, g, \alpha) = m(W, g) - \theta + \alpha(Z)\big(Y - g(Z)\big), \tag{4}$$

where g and α approximate g_0 and α_0. It is worth noting that in our sample selection model, Y is only observed when $S = 1$, so we basically consider $Y = SY$. As shown by Chernozhukov et al. (2022c), evaluating at the true θ_0 yields

$$\mathbb{E}[\psi(W, \theta_0, g, \alpha)] = -\,\mathbb{E}\big[(\alpha(Z) - \alpha_0(Z))(g(Z) - g_0(Z))\big].$$

Thus, the score is doubly robust: its expectation is zero if either $g = g_0$ or $\alpha = \alpha_0$, and errors enter only through their product. Combined with cross-fitting, this property delivers $\sqrt{n}$-consistent inference with flexible first stages (Chernozhukov et al., 2018).

The Riesz formulation also plays an important role for estimation and sensitivity analysis. It casts the problem as learning a weighting function α_0 jointly with g_0, which aligns well with variational, adversarial, and forest-based methods and can improve numerical stability and transparency. By learning the Riesz representer directly rather than relying on plug-in inverse probability weights, this approach can reduce instability when estimated propensities or selection probabilities are small. Most crucially for our case, the same representer-based structure naturally supports the sensitivity analysis in Section 3.3.

3.2. Riesz Representation Approach under Sample Selection

Our goal is to identify the Average Treatment Effect (ATE), $\theta_0 = E[Y(1) - Y(0)]$, in the sample selection model described in Section 2, where non-random treatment assignment occurs simultaneously with non-random sample selection. Under Assumptions 1–4, the ATE admits the following

representation in the *long model* (i. e., in a hypothetical setting where the latent factors A were observed):

$$\theta_0 = \mathbb{E}[m(W, g_0)] = \mathbb{E}[g_0(1, X, A) - g_0(0, X, A)]\,, \tag{5}$$

where $W := (Y, D, S, X, A)$ is the so-called long data vector and $g_0(d, x, a) := \mathbb{E}[Y \mid D = d, S = 1, X = x, A = a]$ is the long regression. Since A is not observed in practice, we are only able to identify the so-called "short" parameter

$$\theta_s = \mathbb{E}[m(W_s, g_s)] = \mathbb{E}[g_s(1, X) - g_s(0, X)]$$

from the observed short data vector $W_s := (Y, D, S, X)$, where $g_s(d, X) = \mathbb{E}[Y|D = d, S = 1, X]$ is the short regression. Since both parameters have a representation of the form $\theta = \mathbb{E}[m(W, g)]$, the Riesz Representation Theorem guarantees the existence of a Riesz representer α, such that $\theta = \mathbb{E}[\alpha(Z)g(Z)]$. The following main theorem of this paper, provides the explicit form of the Riesz representer in sample selection models.

Theorem 1 *Under the Assumptions 1, 2, 3 and 4, the Riesz representers of the long parameter θ_0 and the short parameter θ_s are given by*

$$\alpha_0(w) = \frac{\mathbf{1}\{D = 1\} \cdot S}{p_1(X)\pi_0(1, X, A)} - \frac{\mathbf{1}\{D = 0\} \cdot S}{p_0(X)\pi_0(0, X, A)}$$

and

$$\alpha_s(w) = \frac{\mathbf{1}\{D = 1\} \cdot S}{p_1(X)\pi_s(1, X)} - \frac{\mathbf{1}\{D = 0\} \cdot S}{p_0(X)\pi_s(0, X)},$$

where $p_d(X) := \mathbb{P}(D = d|X)$ is the propensity score for $d \in \{0, 1\}$, $\pi_0(d, X, A) = \mathbb{P}(S = 1|D = d, X, A)$ accounts for selection in the long parameter, and $\pi_s(d, X) = \mathbb{P}(S = 1|D = d, X)$ accounts for selection in the short parameter.

The formal proof is given in Appendix A. Intuitively, the Riesz representer reweights the data to mirror what we would see in a randomized experiment. Weighting by $1/\mathbb{P}(D = d \mid X)$ increases the influence of units with observed characteristics X that are unlikely to receive treatment d. This reweighting aligns the distribution of observed confounders, mimicking the balance achieved through random assignment. Introducing sample selection creates an additional challenge: outcomes are observed only when $S = 1$. To correct for this, we apply a second set of inverse-probability weights based on the likelihood of selection. Since we do not rely on conditional independence of selection in Equation (1) but rather on Assumption 2, conditional independence of selection holds only after controlling for the unobserved variables A. As a result, our correction for selection must also account for these unobservables. Weighting by $1/\pi_0(d, X, A)$ in the long parameter, or $1/\pi_s(d, X)$ in the short parameter, gives more weight to units that were less likely to be selected into the observed sample, thereby restoring representativeness relative to the full population. The distinction between long and short parameters reflects whether the weighting scheme accounts for the unobserved confounders A in the selection process S or not.

3.3. Sensitivity Analysis

With observed data we are only able to identify the short parameter, although we are interested in the long parameter θ_0. The Riesz representer theorem gives us a direct formula for the omitted variable

bias arising from not controlling for A in the selection into observability. Following Chernozhukov et al. (2022a), the difference between the long parameter θ_0 and the short parameter θ_s is given by

$$\theta_0 - \theta_s = \mathbb{E}[(g_0 - g_s)(\alpha_0 - \alpha_s)],$$

which can be interpreted as the covariance between the error parts of g and α. Therefore, the (squared) bias is bounded by

$$|\theta_0 - \theta_s|^2 = \rho^2 B^2 \leq B^2,$$

where $B^2 := \mathbb{E}[(g_0 - g_s)^2]\mathbb{E}[(\alpha_0 - \alpha_s)^2]$ and $\rho^2 := \text{Cor}^2(g_0 - g_s, \alpha_0 - \alpha_s)$. Furthermore, this squared bias bound B^2 has an intuitive decomposition that helps to understand the role of confounding in sample selection models. The squared bias bound B^2 can be decomposed as

$$B^2 = \widetilde{S}^2 C_Y^2 C_S^2,$$

where $\widetilde{S}^2 := \mathbb{E}[(Y - g_s)^2]\mathbb{E}[\alpha_s^2]$, $C_Y^2 := \frac{\mathbb{E}[(g_0 - g_s)^2]}{\mathbb{E}[(Y - g_s)^2]}$ and $C_S^2 := \frac{\mathbb{E}[(\alpha_0 - \alpha_s)^2]}{\mathbb{E}[\alpha_s^2]}$. Therefore, the bound B^2 is the product of $\widetilde{S}^2$, a scaling factor identifiable from observed data, C_Y^2 that measures confounding strength in the outcome equation and C_S^2 that measures confounding strength in the selection equation. For C_Y^2 and C_S^2 researchers need to make informed assumptions about the impact of unobserved confounding. More formally, it holds that

$$C_Y^2 = \frac{\mathbb{E}[(g_0 - g_s)^2]}{\mathbb{E}[(Y - g_s)^2]} = R^2_{Y - g_s \sim g_0 - g_s} = \eta^2_{Y \sim A | D, X, S=1},$$

which measures the proportion of residual outcome variation (variation not explained by observed variables) that can be explained by the latent confounders A. It is by definition $\eta^2_{Y \sim A | D, X, S=1}$, the partial R^2 of Y on the confounder A, after adjusting for D and X, conditional on $S = 1$. Further, it holds

$$C_S^2 = \frac{\mathbb{E}[\alpha_0^2] - \mathbb{E}[\alpha_s^2]}{\mathbb{E}[\alpha_s^2]} = \frac{1 - R^2_{\alpha_0 \sim \alpha_s}}{R^2_{\alpha_0 \sim \alpha_s}}, \quad \text{with} \quad R^2_{\alpha_0 \sim \alpha_s} = \frac{\mathbb{E}[\alpha_s^2]}{\mathbb{E}[\alpha_0^2]}.$$

It is worth noting that $R^2_{\alpha_0 \sim \alpha_s} = \mathbb{E}[\alpha_s^2]/\mathbb{E}[\alpha_0^2]$ measures how much variation in the true Riesz representer α_0 is explained by the short Riesz representer α_s. Therefore, $1 - R^2_{\alpha_0 \sim \alpha_s}$ (bounded between 0 and 1) measures the proportion of variation in α_0 that is explained by the omitted confounder A. While this parameter also admits an interpretation as a gain in precision, we find it more informative to use the following quasi-Gaussian approach for interpretation:

Quasi-Gaussian Selection Sensitivity. In practical applications, it might be difficult to think of plausible values for $1 - R^2_{\alpha_0 \sim \alpha_s}$, a technical and likely unfamiliar parameter. Instead, we find it useful to represent the selection indicator S in a form of a latent index S^* with Gaussian shocks crossing a threshold: Let $S = \mathbf{1}\{S^* > 0\}$ with

$$S^* = h(D, X) - U \quad \text{and} \quad U \mid D, X \sim N(0, 1).$$

This representation does not entail loss of generality [1].We can then model confounding as follows:

$$U = \mu_S A + \sqrt{1 - \mu_S^2}\, \varepsilon_S, \quad \text{with} \quad A, \varepsilon_S \overset{\text{i.i.d.}}{\sim} N(0, 1),$$

1. Given $\pi_s(D, X) = P(S = 1 \mid D, X)$ with $\pi_s(D, X) \in (0, 1)$, we can take $U \sim N(0, 1)$ independent of (D, X) and set $h(D, X) = \Phi^{-1}(\pi_s(D, X))$, so that S has the same conditional distribution as $h(D, X) - U > 0$. This construction only uses the subsequent Gaussian parameterization as an interpretation / calibration device.

independent of (D, X). Thus, μ_S^2 is the R^2 in the regression of the Gaussian shock U on the latent confounder A. By definition, it is also equal to $\eta^2_{S^* \sim A|D,X}$, the nonparametric partial R^2 in the regression of the latent index S^* on A, after nonparametrically partialling out (D, X).

It is therefore easy to interpret. We can also map μ_S^2 to the technical sensitivity parameter as follows. We compute the short selection probability $\pi_s(d, x) = \mathbb{P}(S = 1 \mid D = d, X = x) = \Phi(h(d, x))$, so $h(d, x) = \Phi^{-1}(\pi_s(d, x))$ is identified from the short model, and the long probability is

$$\pi_0(d, x, a) = \mathbb{P}(S = 1 \mid D = d, X = x, A = a) = \Phi\left((h(d, x) - \mu_S a)/\sqrt{1 - \mu_S^2}\right).$$

We show in Appendix C that $\mathbb{E}[\alpha_0^2]$ and $\mathbb{E}[\alpha_s^2]$ can be expressed in terms of these probabilities and can therefore be seen as functions of μ_S^2. We then derive the maps from the interpretable to the technical sensitivity parameters: $\mu_S^2 \mapsto 1 - R^2_{\alpha_0 \sim \alpha_s}(\mu_S^2)$. This yields a one-parameter, probit-scale calibration of selection confounding that is directly compatible with the Riesz-based bias bounds. Note that this does not impose any assumptions on the data, but is rather an interpretation device. While sensitivity analysis maps assumptions about the unobserved confounder A (which might affect both outcome Y and selection S) to potential bias in the ATE estimate θ_s, it does not tell us how plausible those assumptions are. Researchers must therefore make informed judgments about the two partial R^2 measures that capture how strongly A predicts the outcome Y and the selection index S^*. This task can be aided by a benchmarking approach, following Imbens (2003), Altonji et al. (2005), Oster (2019), Cinelli and Hazlett (2020), and Chernozhukov et al. (2022a), which uses the observed influence of specific covariates X_j as a reference point for the potential influence of an unobserved confounder A. We outline this approach in Appendix D.

3.4. Estimation

Since α_s is generally unknown, constructing a feasible estimator based on the orthogonal score (Equation (4)) in the DML framework requires an estimate $\hat{\alpha}$. The traditional method for obtaining $\hat{\alpha}$ is a plug-in approach. While conceptually straightforward, this plug-in approach for estimating the Riesz representer suffers from several drawbacks, particularly in high-dimensional or complex settings. Deriving the analytical form of α_s can be mathematically challenging or even intractable for more complex parameters of interest beyond the standard ATE. The formula for α_s also frequently involves division by estimated probabilities or densities (see, e. g., $\hat{p}_d(X)$ and $\hat{\pi}_s(d, X)$ in Theorem 1). If these estimated quantities are close to zero, the resulting $\hat{\alpha}$ can become extremely large. This occurs when the common support assumption (positivity) is empirically violated in the sample. Such large values can lead to unstable estimates of the target parameter θ_s. Recognizing the limitations of the plug-in method, recent research has focused on methods that estimate the Riesz representer α_s directly, without needing its explicit analytical formula or relying on potentially unstable inverse weighting schemes. Two prominent direct approaches are variational methods (Riesz Regression) and adversarial (minimax) methods (Chernozhukov et al., 2020, 2022c,b).

In this paper, we rely on the ForestRiesz, also developed by Chernozhukov et al. (2022b), that adapts the random forest methodology to estimate the Riesz representer. It estimates the representer by solving a regularized Riesz-regression problem, which is a stable linear inverse problem under weak overlap, in contrast to the plug-in method that is not guaranteed to behave well in the weak-overlap case (for additional simulation studies under week overlap see Appendix F). Within this

framework, the Riesz representer is modeled as locally linear with respect to a pre-specified feature map $a(Z) = \langle r(D, X, S), \beta(X) \rangle$, where $Z = (D, X, S)$, $r(D, X, S)$ represents a smooth feature map (e. g., a polynomial series) and $\beta(X)$ denotes local coefficients that vary with covariates X. The algorithm constrains splits to covariates X exclusively to preserve sufficient variation in the treatment variable D within each node. Chernozhukov et al. (2022b) show that this problem falls in the class of problems defined via solutions to moment equations $m(\cdot) = 0$. Therefore, we can apply the framework of Generalized Random Forests of Athey et al. (2019) to solve this local moment problem via random forests. For each node in the forest, the algorithm computes a Jacobian matrix and a local moment vector

$$J(\text{node}) = \frac{1}{|\text{node}|} \sum_{i \in \text{node}} r(Z_i) r(Z_i)^\top \quad \text{and} \quad M(\text{node}) = \frac{1}{|\text{node}|} \sum_{i \in \text{node}} m(W_i; r).$$

The optimal coefficient vector within each node is given by $\beta(\text{node}) = J(\text{node})^{-1} M(\text{node})$. ForestRiesz grows the forest by recursively splitting nodes based solely on the covariates X. For each candidate split, the two resulting child nodes are evaluated by computing their respective J and M. The splitting rule seeks to maximize the stability-adjusted signal by minimizing the aggregate local Riesz loss:

$$- \sum_{\text{child} \in \{1,2\}} |\text{child}| \; \beta(\text{child})^\top J(\text{child}) \, \beta(\text{child}).$$

This criterion favors splits that yield child nodes where the local moment M is both strong and well-supported by a diverse (i. e., well-spread) feature set, while penalizing splits that produce nodes with nearly singular J. ForestRiesz incorporates multitasking capabilities, wherein the forest simultaneously learns the regression function $\hat{g}$ and the Riesz representer $\hat{\alpha}$ by augmenting the node-splitting criteria with regression-based objectives. The final estimate is given by

$$\hat{\theta}_{\text{DR}} = \mathbb{E}_n \big[\, m(W; \hat{g}) + \hat{\alpha}(Z) \big(Y - \hat{g}(Z) \big) \, \big]$$

or, better yet, its cross-fitted form to avoid overfitting, leveraging Equation (4) as proposed in Chernozhukov et al. (2022b), where $\mathbb{E}_n$ denotes the sample mean.

4. Simulation Study

The finite-sample properties of the proposed ForestRiesz (FR) estimator are assessed with a simulation study. The data-generative process (DGP) follows the conditional missing-at-random (MAR) design outlined in Appendix E of Bia et al. (2024), with pre-treatment covariates X, a selection and treatment indicator $S,\ D \in \{0, 1\}$, error terms $u,\ v$, and w, and an outcome variable Y, that is only observed if $S = 1$:

$$Y_i = \ \theta_0 D_i \ + X_i' \beta_0 + u_i, \quad S_i = \ \mathbf{1}\{D_i \ + \ X_i' \beta_0 \ + \ v_i \ > 0\}, \quad D_i = \mathbf{1}\{X_i' \beta_0 \ + \ w_i > 0\},$$

with $X_i \sim N(0, \sigma_X^2)$, $(u_i, v_i) \sim N(0, \sigma_{u,v}^2)$, and $w_i \sim N(0, 1)$. For MAR to hold, $\sigma_{u,v}^2$ is specified as an identity matrix, implying that conditional on the treatment indicator and covariates none of the unobservables jointly affect the selection and outcome equation.

In the DGP, we set the true ATE to $\theta_0 = 1$. To benchmark the performance of the ForestRiesz, we compare it to an interactive regression model (IRM) (Chernozhukov et al., 2018), which does not

adjust for the sample selection mechanism of the DGP, and to the sample selection model (SSM) by Bia et al. (2024), which uses efficient Neyman-orthogonal score functions within the DML framework to address sample selection. The benchmark estimators are implemented via the *doubleML* package (Bach et al.), using random forests [2] for estimating the nuisance functions and three-fold cross-fitting to prevent overfitting bias.

For the sample sizes $N \in \{1000,\ 4000,\ 16000\}$, Table 1 reports each estimator's average results across 200 Monte Carlo iterations. For each estimator and sample size, it presents the estimate (ATE), the standard error (SE), and the corresponding bias (MAE). Across all sample sizes, the IRM model underestimates θ_0, since it does not account for sample selection. By contrast, both SSM and FR converge to the true $\theta_0 = 1$ when the number of observations increases. Moreover, as standard errors scale with $1/\sqrt{N}$, quadrupling the sample size reduces the standard errors of all estimators by approximately one half.

	IRM			**SSM**			**FR**		
N	ATE	SE	MAE	ATE	SE	MAE	ATE	SE	MAE
1000	0.8017	0.0564	0.1983	1.1046	0.0451	0.1165	1.1306	0.0944	0.1365
4000	0.7457	0.0280	0.2543	1.0863	0.0222	0.0874	1.0677	0.0461	0.0703
16000	0.7046	0.0139	0.2954	1.0621	0.0110	0.0622	1.0349	0.0230	0.0357

Table 1: Average simulation results based on $\theta_0 = 1$ and 200 Monte Carlo iterations.

A more detailed comparison of the SSM and FR simulation results suggests a different bias-variance trade-off. Across all sample sizes, SSM yields smaller standard errors, whereas FR results indicate a faster decline in bias as the sample size increases. It is worth noting that the FR model is used without any tuning, while for the SSM we explored different random forest depths to improve propensity scores estimation and reduce bias. To complement the previously described considerations, Figure 2 in Appendix E.2 presents the distribution of the ATE estimates across all Monte Carlo iterations. Furthermore, Appendix E.3 presents additional results for the SSM estimator, showing that under the Lasso specifications used to learn the nuisance parameters in the score of Bia et al. (2024), the SSM bias declines as expected given the linearity of the DGP.

These considerations highlight the importance of the choice of machine learning methods and hyperparameter tuning in the SSM approach, and more generally within the DML framework (Bach et al. (2024b)), and demonstrate that the FR approach is considerably more robust. A more detailed empirical comparison between the DML-based methods and the Riesz representer approach is left for future research.

5. Application

As an empirical application, we apply our method to estimate the gender wage gap in the U. S. We use data from the 2016 American Community Survey (ACS), which provides a representative 1 % sample of the U. S. population under mandatory participation. Since some respondents do not report their wages, even though they are employed, any gender wage gap analysis based on the ACS data is subject to a sample selection problem. The dataset contains 158 variables for socio-economic characteristics at the individual and the household level, for example referring to education, industry, and occupation. We follow the study of Bach et al. (2024a) and focus on two sub-populations in the

2. For the exact specification of hyperparameters of the random forests and the DML parameters see Appendix E.1.

ACS: respondents with a high school degree and those with a college degree. Our treatment variable D is the gender of a respondent, with $D = 1$ indicating a female respondent. Our outcome variable Y denotes (log) weekly wages (in USD) and the indicator S indicates whether Y is observed (i. e., the respondent has reported her wage). In this context, A is a latent determinant of wage non-reporting that also drives (log) weekly wages after controlling for X. Plausible examples include unobserved disclosure preferences, compensation features (e. g., bonuses or tips), or personality traits (e. g., conscientiousness or neuroticism).

In the high school sub-population, we have 372 728 respondents and in the college sub-population 297 178 individuals. In order to estimate the gender wage gap, we apply the proposed ForestRiesz, where one fits a random forest that jointly learns the Riesz representer α and the regression function g in one step as described in Section 3.4. To demonstrate the relevance of our Riesz representer approach in sample selection models, we compare our estimation results with those obtained from the interactive regression model (IRM) and the SSM approach, both implemented using the *doubleML* package (Bach et al.), as in the simulation study. We apply the three estimators to the high school and college subsamples and report point estimates, standard errors, and p-values. Table 2 presents the estimation results for the college and the high school subsamples. For all three regression models, we find a significant gender wage gap in both subsamples, with a larger gap in the high school subsample than in the college subsample, in line with previous findings in Bach et al. (2024a). Since the estimated wage gap is approximately 3 percentage points larger using the Riesz representer approach compared to IRM, our results suggest that we underestimate the gender wage gap when not controlling for non-reporting respondents. Applying a logit model to the reporting indicator S, we find that never-married female workers with a high university degree (professional degree) have a higher probability of reporting their income than their male counterparts, and that the relationship between experience and reporting also differs between men and women (see Table 9 in Appendix G). Because these covariates are also among the strongest predictors of wages (see Table 10 in Appendix G), estimates of the gender wage gap are subject to selection bias if these patterns are ignored. While the IRM model does not address this issue, both the ForestRiesz (FR) and the SSM approach correct for it by reweighting respondents with a lower probability of wage reporting.

	IRM		**SSM**		**FR**	
	College	High school	College	High school	College	High school
Estimate	-0.0989***	-0.141***	-0.153***	-0.198***	-0.128***	-0.172***
SE	0.003	0.003	0.001	0.001	0.002	0.002
P-value	0.000	0.000	0.000	0.000	0.000	0.000

Table 2: Estimation results for the gender wage gap. Significance: $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.10$.

Next, we conduct a sensitivity analysis to assess the robustness of our estimated treatment effects to unobserved confounding. Using observed covariates X_j as benchmarks, this approach evaluates how influential an unobserved confounder A would need to be to overturn our main findings. We perform this analysis for all covariates and report results for the six most influential covariate groups in the college subsample in Table 11 in Appendix G. For each group j, the table reports the share of additional outcome variation $G_{Y,j}$, selection variation $G_{S,j}$, and their alignment measure ρ_j, detailed in Appendix D. Overall, the results indicate that the estimated gender wage gap in the college subsample is highly robust. Omitting the most influential covariate group, marital status,

changes the ATE estimate by only 0.55 percentage points. Notably, although education explains the largest share of variation in wages and in the Riesz representer (high G_Y and G_S), it has virtually no effect on the estimated gender wage gap (low $\Delta\theta$), reflecting the weak correlation between the residual component of the outcome and Riesz representer models (small $|\rho|$). We further assess robustness through sensitivity analyses based on these benchmarks. First, we construct confidence intervals that account for unobserved confounding as strong as the marital status covariate. Figure 11 in Appendix G shows that even under this conservative scenario, the estimated ATE remains statistically significant. Second, we examine the magnitude of unobserved confounding required to overturn our conclusions. Figure 12 in Appendix G illustrates the potential bias as a function of $C_Y^2 = \eta^2_{Y\sim A|D,X,S=1}$ and $\eta^2_{S^*\sim\ A|D,X}$, assuming the worst-case alignment ($\rho = 1$). The robustness value (RV) for the college subsample is 0.063, implying that an unobserved confounder would need to explain at least $6.3\,\%$ of both residual outcome and selection variation to nullify the estimated effect. This is substantially more than any observed covariate in our data can explain.

6. Conclusion

One main contribution of the paper is a bounds analysis for treatment effects when the traditional sample-selection model's conditional missing-at-random (MAR) assumption fails. Although MAR is widely used, it is often hard to defend in applications. We relax MAR by introducing a latent confounder that affects selection and then derive the Riesz representer for the average treatment effect (ATE), which combines treatment-propensity weighting with selection-probability weighting. Using the resulting Riesz representers for the short and long models, we decompose the omitted-variable bias into three interpretable components. This decomposition yields sharp, distribution-free bounds on the magnitude of bias and provides a practical sensitivity-analysis for MAR violations.

A second contribution is to adapt the ForestRiesz method of Chernozhukov et al. (2022b) to treatment-effect estimation under sample selection. This automatic debiased machine learning approach jointly learns the outcome regression and the Riesz representer, avoiding the numerical instability of plug-in estimators that require direct inversion of estimated probabilities. Our simulations highlight the advantages of the ForestRiesz framework over more standard doubly robust plug-in approaches in finite samples. We illustrate the practical benefits of the method in an application to the U. S. gender wage gap using the American Community Survey. We find that ignoring sample selection leads to an underestimation of the wage gap, driven by systematic gender differences in wage reporting. A benchmarking-based sensitivity analysis indicates that this conclusion is robust.

Overall, our results highlight the importance of explicitly accounting for sample selection, particularly in survey-based studies, and demonstrate that the ForestRiesz estimator offers a robust, interpretable, and computationally attractive approach for causal inference in the presence of selective outcome observability.

Acknowledgments

We acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — Project Number 530859036.

References

Joseph G. Altonji, Todd E. Elder, and Christopher R. Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113(1): 151–184, 2005. doi: 10.1086/426036. URL https://doi.org/10.1086/426036.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019. doi: 10.1214/18-AOS1709. URL https://doi.org/10.1214/18-AOS1709.

Philipp Bach, Victor Chernozhukov, Sven Klaassen, Malte S. Kurz, and Martin Spindler. DoubleML - Double Machine Learning in Python. URL https://github.com/DoubleML/doubleml-for-py.

Philipp Bach, Victor Chernozhukov, and Martin Spindler. Heterogeneity in the us gender wage gap. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187(1):209–230, 2024a. doi: 10.1093/jrsssa/qnad091. URL https://doi.org/10.1093/jrsssa/qnad091.

Philipp Bach, Oliver Schacht, Victor Chernozhukov, Sven Klaassen, and Martin Spindler. Hyperparameter tuning for causal inference with double machine learning: A simulation study. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 1065–1117. PMLR, 01–03 Apr 2024b. URL https://proceedings.mlr.press/v236/bach24a.html.

Philipp Bach, Sven Klaassen, Jannis Kueck, Mara Mattes, and Martin Spindler. Sensitivity analysis for treatment effects in difference-in-differences models using riesz representation, 2025. URL https://arxiv.org/abs/2510.09064.

Michela Bia, Martin Huber, and Lukáš Lafférs. Double machine learning for sample selection models. *Journal of Business & Economic Statistics*, 42(3):958–969, 2024. doi: 10.1080/07350015.2023.2271071. URL https://doi.org/10.1080/07350015.2023.2271071.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009*, 2020. URL https://arxiv.org/abs/2101.00009.

Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, July 2022a. URL http://www.nber.org/papers/w30302.

Victor Chernozhukov, Whitney Newey, Víctor M. Quintas-Martínez, and Vasilis Syrgkanis. RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and

Sivan Sabato, editors, *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3901–3914. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/chernozhukov22a.html.

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022c. doi: https://doi.org/10.3982/ECTA18515. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA18515.

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*, 25(3):576–601, 04 2022d. doi: 10.1093/ectj/utac002. URL https://doi.org/10.1093/ectj/utac002.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 2020. doi: 10.1111/rssb.12348. URL https://doi.org/10.1111/rssb.12348.

Sofiia Dolgikh and Bodan Potanin. Double machine learning for causal inference in a multivariate sample selection model. *arXiv preprint arXiv:2511.12640*, 2025. URL https://arxiv.org/abs/2511.12640.

Jerry A. Hausman and David A. Wise. Attrition bias in experimental and panel data: The gary income maintenance experiment. *Econometrica: Journal of the Econometric Society*, 47(2):455–473, 1979. doi: 10.2307/1914193. URL https://doi.org/10.2307/1914193.

James J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement*, volume 5, pages 475–492. NBER, October 1976. URL http://www.nber.org/chapters/c10491.

James J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, 47:153–161, 1979. doi: 10.2307/1912352. URL https://doi.org/10.2307/1912352.

Miguel A. Hernán, Sonia Hernández-Díaz, and James M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, September 2004. doi: 10.1097/01.ede.0000135174.63482.43. URL https://doi.org/10.1097/01.ede.0000135174.63482.43.

David A. Hirshberg and Stefan Wager. Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206–3227, December 2021. doi: 10.1214/21-AOS2080. URL https://doi.org/10.1214/21-AOS2080.

Ildar A. Ibragimov and Rafail Z. Has'minskii. On the nonparametric estimation of functionals. In J. Hájek, editor, *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, pages pp. 267–281, Prague, Czechoslovakia, 1979. Charles University Press.

Guido W. Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, May 2003. doi: 10.1257/000282803321946921. URL https://doi.org/10.1257/000282803321946921.

B. Ya. Levit. On the efficiency of a class of non-parametric estimates. *Theory of Probability and Its Applications*, 20(4):723–740, 1975. doi: 10.1137/1120081.

Roderick J. A. Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121, 1995. doi: 10.1080/01621459.1995.10476615. URL https://doi.org/10.1080/01621459.1995.10476615.

Whitney K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 62(6):1349–1382, November 1994. doi: 10.2307/2951752. URL https://doi.org/10.2307/2951752.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019. doi: 10.1080/07350015.2016.1227711. URL https://doi.org/10.1080/07350015.2016.1227711.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818. URL https://doi.org/10.1080/01621459.1994.10476818.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350. URL https://doi.org/10.1037/h0037350.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, December 1976. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

Donald B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26, 1977. doi: 10.3102/10769986002001001. URL https://doi.org/10.3102/10769986002001001.

Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980. doi: 10.2307/2287653. URL https://doi.org/10.2307/2287653.

Appendix A. Proof of Theorem 1

We derive the result for the long parameter, as the proof for the short parameter is analogous. We aim to show that

$$\theta_0 = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[m(W, g_0)] = \mathbb{E}[g_0(D, X, A)\alpha_0(W)]$$

with $g_0(d, x, a) := \mathbb{E}[Y \mid D = d, S = 1, X = x, A = a]$ and $m(W, g_0) := g_0(1, X, A) - g_0(0, X, A)$.

Step 1: First, we show that $\theta_0 = \mathbb{E}[m(W, g_0)]$. It suffices to show that, for each $d \in \{0, 1\}$, it holds that

$$\mathbb{E}[g_0(d, X, A)] = \mathbb{E}[Y(d)].$$

Fix $d \in \{0, 1\}$. Then,

$$\begin{aligned}
\mathbb{E}[g_0(d, X, A)] &= \mathbb{E}[\mathbb{E}[Y \mid D = d, S = 1, X, A]] \\
&= \mathbb{E}[\mathbb{E}[Y(d) \mid D = d, S = 1, X, A]] && \text{(Observational Rule)} \\
&= \mathbb{E}[\mathbb{E}[Y(d) \mid D = d, X, A]] && \text{(Assumption 2)} \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[Y(d) \mid D = d, X, A] \mid X]] && \text{(Law of Iterated Expectation)} \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[Y(d) \mid D = d, X, A] \mid D = d, X]] && \text{(Assumption 3)} \\
&= \mathbb{E}[\mathbb{E}[Y(d) \mid D = d, X]] && \text{(Law of Iterated Expectation)} \\
&= \mathbb{E}[\mathbb{E}[Y(d) \mid X]] && \text{(Assumption 1)} \\
&= \mathbb{E}[Y(d)].
\end{aligned}$$

Therefore,

$$\mathbb{E}[m(W, g_0)] = \mathbb{E}[g_0(1, X, A)] - \mathbb{E}[g_0(0, X, A)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \theta_0.$$

Step 2: *Verify the Riesz representer.* Define

$$\alpha_0(W) := \frac{\mathbf{1}\{D = 1\}S}{p_1(X)\pi_0(1, X, A)} - \frac{\mathbf{1}\{D = 0\}S}{p_0(X)\pi_0(0, X, A)}.$$

It holds that

$$\begin{aligned}
\mathbb{E}[\alpha_0(W)\, g_0(D, X, A)] &= \mathbb{E}[\mathbb{E}[\alpha_0(W)\, g_0(D, X, A) \mid X, A]] \\
&= \mathbb{E}\left[g_0(1, X, A)\frac{\mathbb{E}[\mathbf{1}\{D = 1\}S \mid X, A]}{p_1(X)\pi_0(1, X, A)} - g_0(0, X, A)\frac{\mathbb{E}[\mathbf{1}\{D = 0\}S \mid X, A]}{p_0(X)\pi_0(0, X, A)}\right] \\
&= \mathbb{E}[g_0(1, X, A) - g_0(0, X, A)] = \mathbb{E}[m(W, g_0)],
\end{aligned}$$

where we used that

$$\begin{aligned}
\mathbb{E}[\mathbf{1}\{D = d\}S \mid X, A] &= \mathbb{P}(D = d, S = 1 \mid X, A) \\
&= \mathbb{P}(D = d \mid X, A)\, \mathbb{P}(S = 1 \mid D = d, X, A) = p_d(X)\, \pi_0(d, X, A),
\end{aligned}$$

with $\mathbb{P}(D = d \mid X, A) = p_d(X)$ by Assumption 3.

Appendix B. Omitted Variable Bias in Sample Selection Models

Here, we apply the framework of Chernozhukov et al. (2022a) to derive the omitted variable bias in the sample selection model with confounding in selection. Let θ_0 denote the long parameter and θ_s the short parameter,

$$\theta_0 = \mathbb{E}[m(W, g_0)] \quad and \quad \theta_s = \mathbb{E}[m(W_s, g_s)],$$

where g_0 and g_s are the long and short outcome regressions defined in the main text. Let α_0 and α_s be the corresponding long and short Riesz representers. The omitted variable bias (OVB) admits the representation

$$\theta_0 - \theta_s \;=\; \mathbb{E}\big[(g_0 - g_s)(\alpha_0 - \alpha_s)\big]. \tag{6}$$

As described in the main text, it holds that

$$|\theta_0 - \theta_s|^2 = \rho^2 B^2 \leq B^2 \tag{7}$$

with

$$B^2 = \widetilde{S}^2 \, C_Y^2 \, C_S^2, \tag{8}$$

where $\widetilde{S}^2$ is identified from the observed data, while C_Y^2 and C_S^2 summarize the strength of omitted-variable effects in the outcome and selection components, respectively. In particular,

$$C_Y^2 = \frac{\mathbb{E}[(g_0 - g_s)^2]}{\mathbb{E}[(Y - g_s)^2]} = R^2_{Y - g_s \sim g_0 - g_s} = \eta^2_{Y \sim A | D, X, S=1}$$

is the fraction of residual outcome variation (after controlling for observed covariates) that is explained by the omitted confounder through the long regression.

Next, we consider the sensitivity parameter C_S^2 in more detail. Let $\mathcal{A}_s$ be the closed linear subspace of $L^2(\mathbb{P})$ consisting of square-integrable functions measurable with respect to the short information set (the observed variables in W_s). Since the long functional restricted to $\mathcal{A}_s$ has the Riesz representer α_s, we have

$$\mathbb{E}[(\alpha_0 - \alpha_s)a] = 0 \quad \text{for all } a \in \mathcal{A}_s,$$

so α_s is the L^2-projection of α_0 onto $\mathcal{A}_s$. Taking $a = \alpha_s$ yields $\mathbb{E}[\alpha_0 \alpha_s] = \mathbb{E}[\alpha_s^2]$, and hence

$$\mathbb{E}[(\alpha_0 - \alpha_s)^2] = \mathbb{E}[\alpha_0^2] - \mathbb{E}[\alpha_s^2]. \tag{9}$$

Therefore,

$$C_S^2 = \frac{\mathbb{E}[\alpha_0^2] - \mathbb{E}[\alpha_s^2]}{\mathbb{E}[\alpha_s^2]} = \frac{1 - R^2_{\alpha_0 \sim \alpha_s}}{R^2_{\alpha_0 \sim \alpha_s}}, \qquad \text{with} \qquad R^2_{\alpha_0 \sim \alpha_s} := \frac{\mathbb{E}[\alpha_s^2]}{\mathbb{E}[\alpha_0^2]}. \tag{10}$$

The quantity $1 - R^2_{\alpha_0 \sim \alpha_s}$ measures the share of variation in the long representer that is not captured by the short representer. Next, we consider the closed-form expressions for the Riesz representers.

Closed-form expressions for $\mathbb{E}[\alpha_0^2]$ and $\mathbb{E}[\alpha_s^2]$

The long Riesz representer is given by

$$\alpha_0(W) := \frac{\mathbf{1}\{D=1\}S}{p_1(X)\pi_0(1,X,A)} - \frac{\mathbf{1}\{D=0\}S}{p_0(X)\pi_0(0,X,A)},$$

and the short Riesz representer by

$$\alpha_s(W_s) := \frac{\mathbf{1}\{D=1\}S}{p_1(X)\pi_s(1,X)} - \frac{\mathbf{1}\{D=0\}S}{p_0(X)\pi_s(0,X)}.$$

Because $\mathbf{1}\{D=1\}\mathbf{1}\{D=0\}=0$, the cross term vanishes and therefore

$$\mathbb{E}[\alpha_0^2] = \mathbb{E}\left[\left(\tfrac{\mathbf{1}\{D=1\}S}{p_1(X)\pi_0(1,X,A)}\right)^2\right] + \mathbb{E}\left[\left(\tfrac{\mathbf{1}\{D=0\}S}{p_0(X)\pi_0(0,X,A)}\right)^2\right].$$

For $d \in \{0,1\}$, we have

$$\begin{aligned}\mathbb{E}\left[\left(\tfrac{\mathbf{1}\{D=d\}S}{p_d(X)\pi_0(d,X,A)}\right)^2 \,\middle|\, X,A\right] &= \frac{\mathbb{E}[\mathbf{1}\{D=d\}S \mid X,A]}{p_d(X)^2\pi_0(d,X,A)^2} \\ &= \frac{p_d(X)\pi_0(d,X,A)}{p_d(X)^2\pi_0(d,X,A)^2} = \frac{1}{p_d(X)\pi_0(d,X,A)}.\end{aligned}$$

Hence,

$$\mathbb{E}[\alpha_0^2] = \mathbb{E}\left[\frac{1}{p_1(X)\pi_0(1,X,A)} + \frac{1}{p_0(X)\pi_0(0,X,A)}\right]. \tag{11}$$

Similarly, we can show that

$$\mathbb{E}[\alpha_s^2] = \mathbb{E}\left[\frac{1}{p_1(X)\pi_s(1,X)} + \frac{1}{p_0(X)\pi_s(0,X)}\right]. \tag{12}$$

Combining the Equations (9)–(12) gives

$$C_S^2 = \frac{\mathbb{E}\left[\frac{1}{p_1(X)\pi_0(1,X,A)} + \frac{1}{p_0(X)\pi_0(0,X,A)}\right] - \mathbb{E}\left[\frac{1}{p_1(X)\pi_s(1,X)} + \frac{1}{p_0(X)\pi_s(0,X)}\right]}{\mathbb{E}\left[\frac{1}{p_1(X)\pi_s(1,X)} + \frac{1}{p_0(X)\pi_s(0,X)}\right]}. \tag{13}$$

The terms $1/(p_d(X)\pi(\cdot))$ grow when either the treatment propensity $p_d(X)$ or the selection probability $\pi(\cdot)$ is small. Thus, $\mathbb{E}[\alpha_0^2]$ and $\mathbb{E}[\alpha_s^2]$ summarize the overlap and selection difficulty through an average inverse-probability scale. Consequently, the sensitivity parameter C_S^2 in Equation (13) measures how much the representer varies when the selection model does or does not depend on the unobserved confounder A, and it can be interpreted as the gain in precision from observing A. The Riesz Representer Framework requires that $\mathbb{E}[\alpha_0^2] < \infty$ and $\mathbb{E}[\alpha_s^2] < \infty$. A convenient sufficient condition is

$$\mathbb{E}\left[\frac{1}{p_1(X)\pi_0(1,X,A)} + \frac{1}{p_0(X)\pi_0(0,X,A)}\right] < \infty,\ \mathbb{E}\left[\frac{1}{p_1(X)\pi_s(1,X)} + \frac{1}{p_0(X)\pi_s(0,X)}\right] < \infty,$$

which we refer to as a *weak overlap* condition as stated in Assumption 4.

Appendix C. Quasi-Gaussian Latent-Index Model for Selection

This section provides an interpretable calibration of C_S^2 using a probit-style latent-index model for selection. The model serves purely as a calibration device and is not required for the identification results presented in the main text.

C.1. Latent-Index Specification

Assume the long selection mechanism admits the representation

$$S = \mathbf{1}\{S^* > 0\}, \qquad S^* = h(D, X) - U, \qquad U = \mu_S A + \sqrt{1 - \mu_S^2}\, \varepsilon_S,$$

where $A, \varepsilon_S \overset{i.i.d.}{\sim} N(0,1)$ and $(A, \varepsilon_S) \perp (D, X)$. Then,

$$\pi_s(d,x) = \mathbb{P}(S = 1 \mid D = d, X = x) = \Phi(h(d,x)), \qquad h(d,x) = \Phi^{-1}(\pi_s(d,x)),$$

and

$$\pi_0(d,x,a) = \mathbb{P}(S = 1 \mid D = d, X = x, A = a) = \Phi\left(\frac{h(d,x) - \mu_S a}{\sqrt{1-\mu_S^2}}\right).$$

The scalar $\mu_S^2 \in [0,1)$ is the latent partial R^2 of A in the selection index, that is:

$$\mu_S^2 = R^2_{S^* \sim A | D, X}.$$

Under the normalization $\mathrm{Var}(U \mid D, X) = 1$, we have $\mathrm{Var}(S^* \mid D, X) = 1$ and

$$\mathrm{Var}(\mathbb{E}[S^* \mid D, X, A] \mid D, X) = \mathrm{Var}(\mu_S A) = \mu_S^2.$$

C.2. Mapping μ_S^2 to C_S^2 and $R^2_{\alpha_0 \sim \alpha_s}$

Given (p_d, π_s) and a choice of $\mu_S^2 \in [0,1)$, define $h(d,x) = \Phi^{-1}(\pi_s(d,x))$ and

$$\pi_0(d,x,a;\mu_S^2) := \Phi\left(\frac{h(d,x) - \sqrt{\mu_S^2}\, a}{\sqrt{1-\mu_S^2}}\right).$$

Then, we can express the Riesz representer α_0 as a function of μ_S^2:

$$\mathbb{E}[\alpha_0^2(\mu_S^2)] = \mathbb{E}\left[\frac{1}{p_1(X)\pi_0(1,X,A;\mu_S^2)} + \frac{1}{p_0(X)\pi_0(0,X,A;\mu_S^2)}\right].$$

The resulting calibration curve is given by

$$C_S^2(\mu_S^2) = \frac{\mathbb{E}[\alpha_0^2(\mu_S^2)] - \mathbb{E}[\alpha_s^2]}{\mathbb{E}[\alpha_s^2]} = \frac{\mathbb{E}[\alpha_0^2(\mu_S^2)]}{\mathbb{E}[\alpha_s^2]} - 1.$$

Equivalently,

$$R^2_{\alpha_0 \sim \alpha_s}(\mu_S^2) = \frac{\mathbb{E}[\alpha_s^2]}{\mathbb{E}[\alpha_0^2(\mu_S^2)]}, \quad \text{and} \quad C_S^2(\mu_S^2) = \frac{1 - R^2_{\alpha_0 \sim \alpha_s}(\mu_S^2)}{R^2_{\alpha_0 \sim \alpha_s}(\mu_S^2)}.$$

C.3. Practical Computation

Let $\widehat{p}_d(X_i)$ and $\widehat{\pi}_s(d, X_i)$ be estimates from the observed data. For a grid of μ_S^2 values, we perform the following steps:

1. Compute $\widehat{h}(d, X_i) = \Phi^{-1}(\widehat{\pi}_s(d, X_i))$.
2. Draw $A_i^{(b)} \sim N(0, 1)$ independently for $b = 1, \ldots, B$, and compute
$$\widehat{\pi}_0^{(b)}(d, X_i) = \Phi\left(\frac{\widehat{h}(d, X_i) - \sqrt{\mu_S^2}\, A_i^{(b)}}{\sqrt{1 - \mu_S^2}}\right).$$
3. Approximate $\mathbb{E}[\alpha_0^2(\mu_S^2)]$ and $\mathbb{E}[\alpha_s^2]$ by
$$\widehat{\mathbb{E}}[\alpha_0^2(\mu_S^2)] = \frac{1}{nB}\sum_{i=1}^{n}\sum_{b=1}^{B}\left(\frac{1}{\widehat{p}_1(X_i)\widehat{\pi}_0^{(b)}(1, X_i)} + \frac{1}{\widehat{p}_0(X_i)\widehat{\pi}_0^{(b)}(0, X_i)}\right)$$
and
$$\widehat{\mathbb{E}}[\alpha_s^2] = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\widehat{p}_1(X_i)\widehat{\pi}_s(1, X_i)} + \frac{1}{\widehat{p}_0(X_i)\widehat{\pi}_s(0, X_i)}\right).$$
4. Report $\widehat{C}_S^2(\mu_S^2) = \widehat{\mathbb{E}}[\alpha_0^2(\mu_S^2)]/\widehat{\mathbb{E}}[\alpha_s^2] - 1$, or $\widehat{R}^2_{\alpha_0 \sim \alpha_s}(\mu_S^2) = \widehat{\mathbb{E}}[\alpha_s^2]/\widehat{\mathbb{E}}[\alpha_0^2(\mu_S^2)]$.

The following figure provides a graphical illustration of the computation:

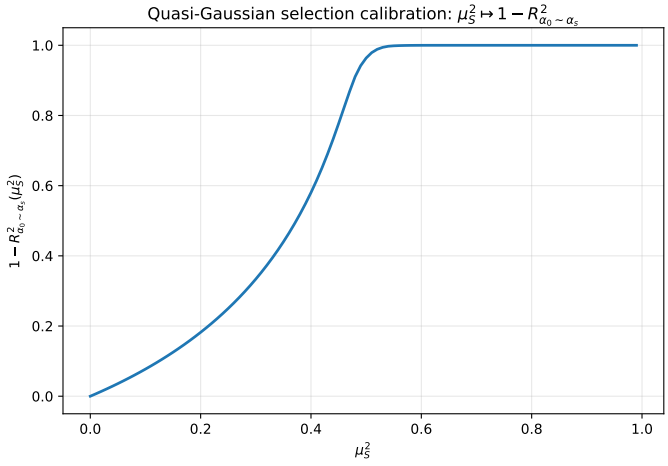


Figure 1: Quasi-Gaussian calibration curve in a synthetic example, where $A \sim N(0, 1)$. The horizontal axis shows values of the classical interpretable parameter, and the vertical axis shows values of the implied technical parameter. Note that this is a rather conservative model, since once $\mu_S^2 \geq 1/2$ the implied bias bound becomes infinite.

Appendix D. Benchmarking Sensitivity to Unobserved Confounding

Relying on benchmarking, we measure how much a specific observed variable X_j actually matters in our data by looking at its influence in four key areas. Let g_s and α_s be the outcome model and the Riesz representer using all covariates X, and let $g_{s,-j}$ and $\alpha_{s,-j}$ be the versions omitting X_j. We consider four quantities to measure the impact of the omitted variable X_j:

1. **Outcome Prediction:** We measure X_j's impact on predicting the outcome Y (within the selected sample, $S = 1$) by calculating the increase in R-squared ($\Delta\eta^2_{Y\sim X_j|D,X_{-j},S=1} := \eta^2_{Y\sim D,X,S=1} - \eta^2_{Y\sim D,X_{-j},S=1}$) when X_j is added to the model. This shows how much X_j improves outcome prediction beyond other variables.

2. **Selection Weights:** We measure X_j's impact on the statistical weights α_s used for correction by calculating the relative change in the weights' overall size ($1 - R^2_{\alpha_s\sim\alpha_{s,-j}} := (\mathbb{E}[\alpha_s^2] - \mathbb{E}[\alpha_{s,-j}^2])/\mathbb{E}[\alpha_s^2]$) when X_j is included. This shows how much X_j changes the necessary adjustment for selection and treatment assignment.

3. **ATE Estimate:** We measure X_j's direct impact on the final result by calculating the change in the ATE estimate ($\Delta\theta_{s,j} := \theta_{s,-j} - \theta_s$) when X_j is included versus excluded as a control variable. This shows how sensitive the estimated ATE is to controlling for X_j.

4. **Alignment of Effects:** We measure whether X_j's effects on the outcome and selection weights work together or against each other by calculating the correlation ($\rho_j := \text{Cor}(g_{s,-j} - g_s, \alpha_s - \alpha_{s,-j})$) between the changes they cause when X_j is removed.

Then, we calculate the following three metrics for X_j to define benchmark values for the sensitivity parameters:

1. **Outcome Gain Metric ($G_{Y,j}$):** This serves as a benchmark for how much A might explain the remaining variance in the outcome Y (after accounting for $D, S = 1$, and X). Hence, it is a proxy for the sensitivity parameter $C_Y^2 = \eta^2_{Y\sim A|D,X,S=1}$, the partial R^2 of Y on the confounder A. The assumption is that A's relative contribution to explaining residual outcome variance is similar to X_j's:

$$G_{Y,j} := \frac{\Delta\eta^2_{Y\sim X_j|D,X_{-j},S=1}}{1 - \eta^2_{Y\sim D,X,S=1}} \approx C_Y^2 = \eta^2_{Y\sim A|D,X,S=1}.$$

   Interpretation: If X_j explains, say, 5 % of the outcome variance that was previously unexplained by $D, S = 1$, and X_{-j} (resulting in $G_{Y,j} = 0.05$), this sets a benchmark. We can then ask: "Is it plausible that the unobserved confounder A explains more than 5 % of the residual outcome variance?" This directly informs the choice of C_Y^2 in the sensitivity analysis.

2. **Selection / Representer Gain Metric ($G_{S,j}$):** This serves as a benchmark for A's association with the selection mechanism, captured by the sensitivity parameter $C_S^2 = (1 - R^2_{\alpha_0\sim\alpha_s})/R^2_{\alpha_0\sim\alpha_s}$ or $1 - R^2_{\alpha_0\sim\alpha_s}$, respectively. Therefore, we link the relative change in the Riesz representer due to A to the change in the Riesz representer due to the observed X_j:

$$G_{S,j} := 1 - R^2_{\alpha_s\sim\alpha_{s,-j}} \approx 1 - R^2_{\alpha_0\sim\alpha_s}.$$

Interpretation: $G_{S,j}$ quantifies how strongly X_j influences the selection mechanism (encoded in α_s), setting a benchmark for the magnitude of A's impact. Higher $G_{S,j}$ values imply a higher threshold for A's assumed effect.

3. **Correlation / Degree of Adversity Metric (ρ_j):** This metric captures how aligned the confounding effects of X_j are on the outcome and selection mechanism (via the RR). It measures the correlation between the change in the outcome model g_s and the change in the Riesz representer α_s when X_j is removed:

$$\rho_j := \text{Cor}(g_{s,-j} - g_s, \alpha_s - \alpha_{s,-j}).$$

Interpretation: ρ_j reflects alignment of X_j's confounding effect. A value close to +1 or -1 indicates that X_j influences both the outcome prediction (within the selected sample) and the selection mechanism representation α_s in a similar way, leading to a larger change in the ATE estimate (larger $\Delta\theta_{s,j}$). We can compare the assumed ρ for A against the observed ρ_j for plausible observed confounders X_j.

Calculating $G_{Y,j}$, $G_{S,j}$, and ρ_j for one or more carefully chosen covariates X_j provides concrete reference points. These points correspond directly to values used in the sensitivity analysis (C_Y^2, C_S^2 and ρ). They help evaluate whether overturning the study's main conclusions would require the unobserved confounder A to be substantially more influential (in terms of outcome variance explained, impact on the selection mechanism's RR structure, or correlation/adversity) than key observed covariates like X_j.

Appendix E. Additional Material for the Simulation Study

E.1. Computational Details

Parameters - scikit-learn	Parameters - doubleML
RandomForest classes	IRM and SSM
n_estimators = 500	n_folds= 3, n_rep= 1
max_depth = 20	SSM
min_samples_leaf = 5	score = 'missing-at-random'
max_features ='sqrt'	normalize_ipw = True

Table 3: This table reports the final hyperparameter set up used for the *RandomForestRegressor* and *RandomForestClassifier* classes from *scikit-learn* (Pedregosa et al., 2011), as well as the settings for the estimator classes *DoubleMLIRM* and *DoubleMLSSM* from the *doubleML* (Bach et al.) Python package. Parameters not reported are kept at their default values.

E.2. Additional Simulation Results: ATE Estimates

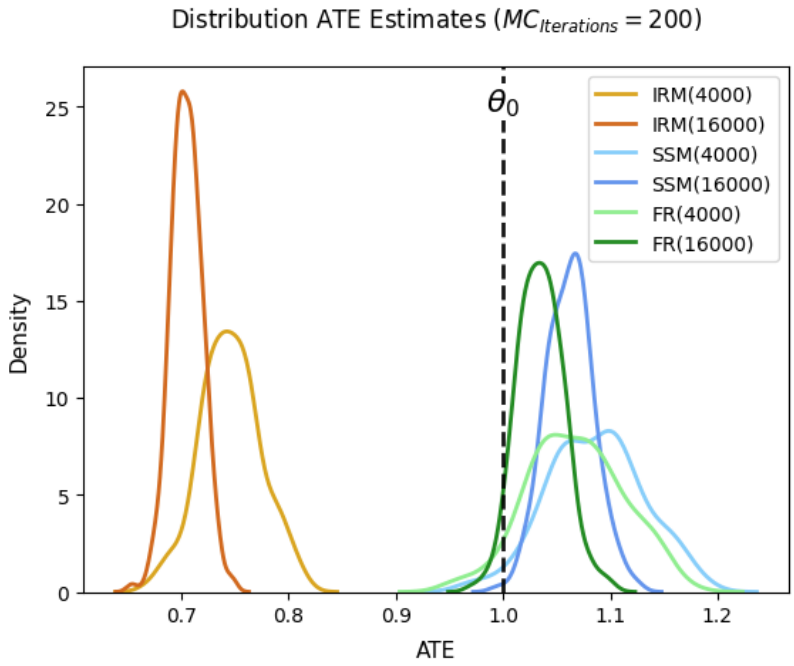


Figure 2: This figure displays the distribution of ATE estimates based on $\theta_0 = 1$ and 200 Monte Carlo iterations. It illustrates that as the sample size grows, the IRM suffers from an increasing downward bias, while both the SSM and FR converge to the simulated ATE.

E.3. Additional Simulation Results: SSM

	Lasso/Logistic			RandomForest		
N	ATE	SE	MAE	ATE	SE	MAE
1000	1.0511	0.0460	0.0863	1.1228	0.0450	0.1287
4000	1.0254	0.0222	0.0393	1.0895	0.0222	0.0901
16000	1.0123	0.0111	0.0205	1.0653	0.0110	0.0653

Table 4: This table presents the average SSM simulation results based on $\theta_0 = 1$ and 200 Monte Carlo replications. It demonstrates that when nuisance functions in the SSM are estimated using Lasso specifications, as in Bia et al. (2024), the bias declines as expected, highlighting the importance of proper hyperparameter tuning when applying random forest learners to estimate the true ATE within the SSM framework.

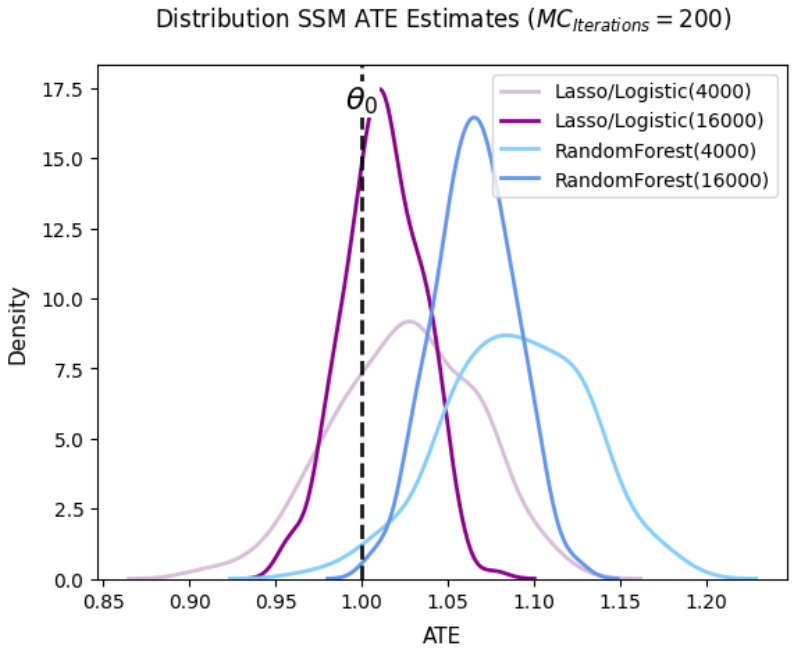


Figure 3: This figure displays the distribution of SSM ATE estimates based on $\theta_0 = 1$ and 200 Monte Carlo iterations. Given the linearity of the DGP, it illustrates that the Lasso/Logistic specification of Bia et al. (2024) converges faster to the true ATE than the random forest specification with $max_depth = 20$.

Table 5: Hyperparameter and Parameter Setup - SSM Simulation

Parameters - scikit-learn	**Parameters - doubleML**
RandomForest classes	SSM
n_estimators = 500	n_folds= 3, n_rep= 1
max_depth = 20	score = 'missing-at-random'
min_samples_leaf = 5	normalize_ipw = True
max_features = 'sqrt'	
LassoCV classes	
max_iter = 50 000	
cv = 3	
LogisticRegressionCV classes	
max_iter = 50 000	
penalty = 'l1'	
solver = 'liblinear'	

Note: This table reports the hyperparameter set up for the *RandomForestRegressor* and *RandomForestClassifier* classes, as well as the *LassoCV* and *LogisticRegressionCV* classes from *scikit-learn* (Pedregosa et al., 2011), as used within the *DoubleMLSSM* framework of the *doubleML* (Bach et al.) Python package, as well as the *doubleML* parameters. These are used to construct the linear and random forest SSM plug-in estimates of the Riesz representer under the original data-generative process. Parameters not reported are kept at their default values.

Appendix F. Weak Overlap Simulation

We build on the data-generative process (DGP) described in Section 4 and induce weak overlap through a perturbation of the covariate distribution. Specifically, we increase the coefficient of the first covariate to $\beta_1 = 1$ and inflate its variance by rescaling it as $2X_{i,1}$ after drawing $X_i \sim N(0, \sigma_X^2)$ during data generation [3]. These modifications increase the variance of X_1 by a factor of four, while preserving the correlation structure of the other covariates. Compared to the original DGP, the weak design thus increases the dispersion of $X_i'\beta_0$ through an amplified contribution of $X_{i,1}$, driven by both its larger coefficient and its higher variance. Thus, as propensity scores approach zero or one for large absolute values of X_1, treatment assignment and selection become nearly deterministic in the tails of the covariate distribution. Figure 4 compares the treatment propensity score $P(D = 1|X)$ and selection probability $P(S = 1 \mid D, X)$ as functions of the covariate X_1 under the original and weak-overlap design.

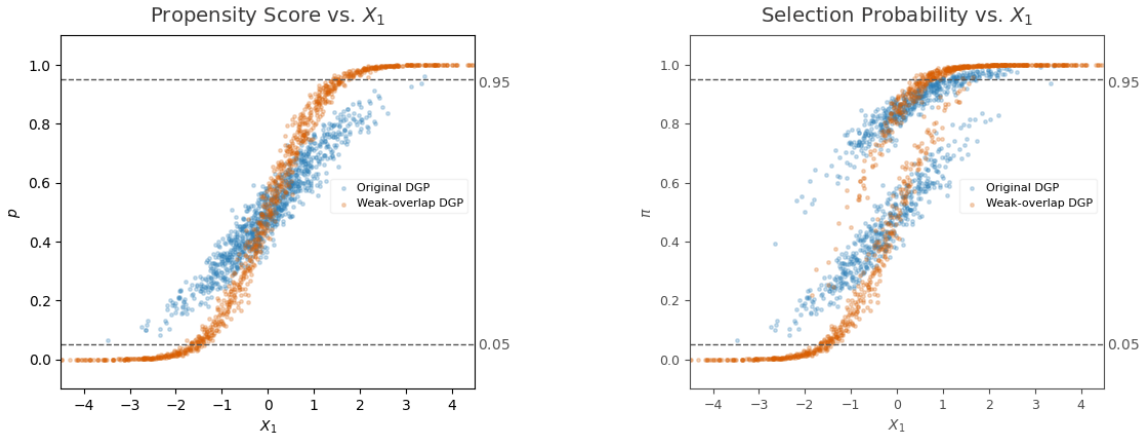


Figure 4: Treatment propensity score and selection probability under the original and weak-overlap design. Oracle treatment propensity scores (left panel) and selection probabilities (right panel) as functions of the covariate X_1 for a sample of size $n = 1\,000$ from the first Monte Carlo iteration in the $p = 100$ covariate setting. Dashed lines indicate regions with near-deterministic treatment assignment and selection.

Under the original DGP, both functions vary smoothly over the support of X_1, with most observations lying within the interval $[0.05, 0.95]$, indicating substantial overlap. In contrast when generating weak-overlap data, both functions become steeper in X_1. Observations with large absolute realizations of X_1 show probabilities close to zero and one. Thus, the weak-overlap adaption shifts probability mass toward the boundaries of the unit interval, with many observations exceeding 0.95 or falling below 0.05, indicating near-deterministic treatment assignment and selection in the tails of the covariate distribution. Such a pattern is consistent with near-violations of the overlap condition and induces an inverse weighting problem in the tails. Estimators relying on propensity scores plug-ins, thus, should be affected by some extreme weights, leading to finite-sample instability.

3. In the original data-generative process $\beta_j = \frac{0.4}{j^2}$, $\forall\, j \in \{1, \ldots, p\}$.

In line with the rest of the paper, and to ensure a fair comparison, we evaluate the stability and performance of the ForestRiesz (FR) estimator in this weak overlap setting relative to two implementations of the sample selection model estimated via double machine learning: a Lasso/Logistic specification (SSM-Lin) and a random forest specification (SSM-RF) for SSM-DML. Both SSM-DML variants are based on the efficient Neyman-orthogonal score derived in Bia et al. (2024), are implemented using the *DoubleML* package (Bach et al.), and rely on three-fold cross-fitting for nuisance function estimation [4]. Moreover, we analyze finite-sample behavior in two covariate settings: $p = 100$ and $p = 400$. For the $p = 100$ specification, results are based on 200 Monte Carlo iterations, whereas for $p = 400$, we report results based on 50 Monte Carlo iterations due to the substantially higher computational cost induced by the expanded covariate space.

The resulting SSM-DML estimates of the treatment and selection propensity scores are then plugged into the analytical Riesz formula, yielding plug-in estimates of the Riesz representer. As a benchmark for these plug-in estimates, we construct the 'Oracle plug-in' Riesz representer based on the generated population data. Specifically, we compute the Riesz representer using the population treatment and selection probabilities, which follow a probit specification consistent with the data-generative process. This oracle Riesz representer serves as a reference for the plug-in SSM-DML estimates of the Riesz representer. It is not intended as a benchmark for the ForestRiesz estimator, which learns the representer directly. To prevent numerical errors arising from near-zero propensity score estimates, all estimated propensity scores are clipped at $\varepsilon_p = \varepsilon_\pi = 1e - 10$. This adjustment serves purely as a numerical safeguard, ensuring that the Riesz representer estimates can be computed.

Figure 5 and Figure 6 highlight the differences between estimates of the Riesz representer obtained using the SSM-DML plug-in methods and those obtained directly via the ForestRiesz approach under weak overlap, for settings with $p = 100$ and $p = 400$ covariates, respectively. While the histograms indicate that across sample sizes all methods concentrate mass near zero, the plug-in estimators exhibit heavier tails, reflecting the persistence of some sizable estimated weights [5]. In contrast, the ForestRiesz estimator produces a more concentrated distribution with substantially reduced tail mass, consistent with enhanced stability under weak-overlap.

To complement the previous finding, Figure 7 and Figure 8 illustrate the concentration of squared Riesz representer estimates across approaches and sample sizes for the $p = 100$ and $p = 400$ covariate setting. For the plug-in SSM-DML estimators, a small fraction of the estimates accounts for a large share of total $\hat{\alpha}^2$, as reflected by the steep initial increase of the concentration curves among all sample sizes. In the setting with $p = 400$ and $n = 1\,000$, the SSM-Lin plug-in estimates exhibit an even higher degree of concentration than the oracle plug-in benchmark. In contrast to the SSM-DML plug-in estimates, the ForestRiesz estimator yields flatter concentration curves, particularly in smaller samples, indicating more evenly distributed weights across observations.

Table 7 reports summary statistics of the estimated Riesz representers under weak overlap, including second moments, maximum absolute values, and upper quantiles of $|\hat{\alpha}|$ for the $p = 100$ and $p = 400$ covariate setting. Across all sample sizes the plug-in SSM-DML estimator exhibit substantially larger second moments and heavier tails, indicating the presence of extreme weights. In contrast, the Forest Riesz estimator yields markedly smaller second moments and tails, reflecting

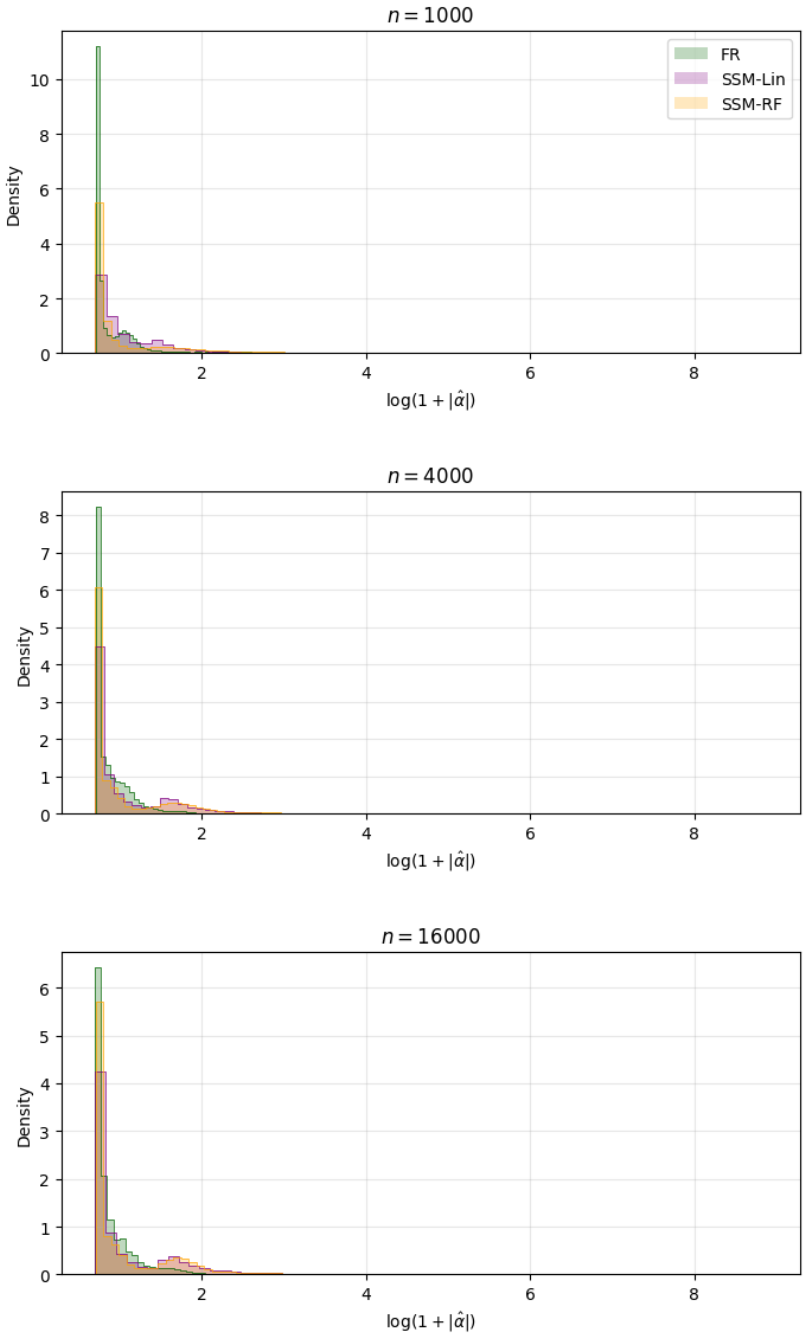
4. The exact hyperparameter specifications for each model, as well as the DML parameters, are reported in Table 8 at the end of this appendix section.
5. Table 6 illustrates, using the $n = 1\,000$ and $p = 100$ setting, that the estimated SSM plug-in Riesz representers can substantially exceed their oracle plug-in benchmark values at the observational level.

a more concentrated distribution. For the $p = 400$ and $n = 1\,000$ setting in Table 7, the SSM-Lin plug-in estimator exhibits substantially larger upper-tail realizations than its oracle plug-in benchmark. This is reflected in the second moments and maximum absolute values of the estimated SSM-Lin Riesz representer relative to the oracle plug-in. Specifically, while the oracle plug-in Riesz representer has a second moment of 298.273 and a maximum absolute value of 1744.652, the linear SSM-DML approach yields 141186.889 and 39584.227, respectively. This, corresponds to increases by factors of approximately 473 for the second moment and 23 for the maximum absolute value. Overall, these findings are consistent with the visual evidence from the distribution and concentration plots and highlight the stability of the ForestRiesz estimator.

While DML-SSM plug-in estimators in general seem to attenuate extreme weights indirectly via nuisance function estimation since cross-fitting reduces the problem of overfitting, ForestRiesz effectively mitigates the influence of extreme observations and yields more balanced weights by direct estimation of the Riesz representer.

Figure 9 and Figure 10 indicate how this affects the target parameter estimation. The linear plug-in SSM estimator is closest to the population parameter value $\theta_0 = 1$ across all sample sizes, suggesting a comparatively low bias. However, in smaller samples ($n = 1\,000$), both plug-in estimators exhibit larger dispersion than ForestRiesz, reflecting their instability due to the variability in the underlying weights. In contrast, the ForestRiesz estimator yields a tighter distribution, particularly in the smallest sample, indicating stability. It, however, exhibits a larger upward bias relative to the linear SSM approach across all sample sizes. As the DGP itself is linear, that SSM-Lin performs best in terms of bias is consistent with expectations. The SSM-RF estimator, by comparison, shows both a larger upward bias and, except for $n = 16\,000$, a larger dispersion than ForestRiesz. Overall, the results indicate that under weak overlap in the linear DGP, SSM-Lin performs best in terms of bias, whereas ForestRiesz performs best in terms of stability.

Figure 5: Histograms of $\log(1 + |\hat{\alpha}|)$ | $p = 100$ | $\text{MC}_{\text{Iterations}} = 200$



Note: Distributions of the estimated Riesz representer based on $p = 100$ and 200 Monte Carlo iterations. The figure displays histograms of $\log(1 + |\hat{\alpha}|)$ for different sample sizes and estimation methods under weak overlap.

Table 6: Largest 10 observation-level differences (Δ) between estimated plug-in Riesz representer and oracle plug-in (based on $n = 1\,000$, $p = 100$ and $\text{MC}_{\text{Iterations}} = 200$)

a) SSM-Lin exceeds Oracle plug-in Riesz representer

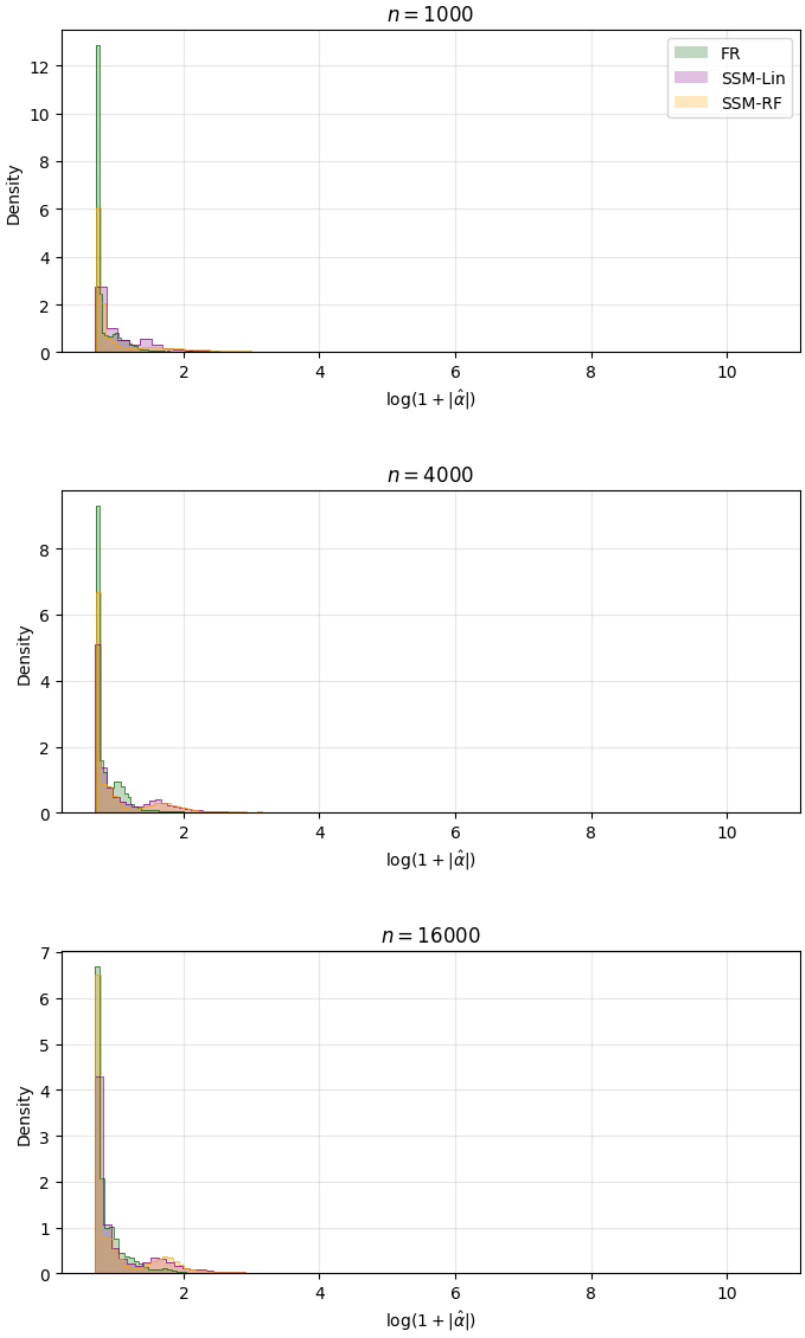
obs_{id}	Oracle plug-in	SSM-Lin	Δ
88432	172.148	7175.331	7003.183
21990	-6551.717	-100.160	6451.556
14527	-4758.055	-170.945	4587.110
48975	-2998.625	-83.918	2914.707
18862	2345.382	4120.301	1774.919
46065	-1414.793	-7.841	1406.953
55290	-1155.834	-46.323	1109.511
74265	-908.107	-87.880	820.228
108451	-818.791	-7.832	810.959
8093	-676.833	-7.080	669.753

b) SSM-RF exceeds Oracle plug-in Riesz representer

obs_{id}	Oracle plug-in	SSM-RF	Δ
21990	-6551.717	-51.468	6500.248
14527	-4758.055	-25.591	4732.464
48975	-2998.625	-18.913	2979.712
46065	-1414.793	-32.831	1381.962
55290	-1155.834	-75.814	1080.020
74265	-908.107	-22.322	885.785
108451	-818.791	-15.787	803.004
8093	-676.833	-11.052	665.781
45514	-626.920	-20.801	606.119
81621	-625.563	-36.015	589.547

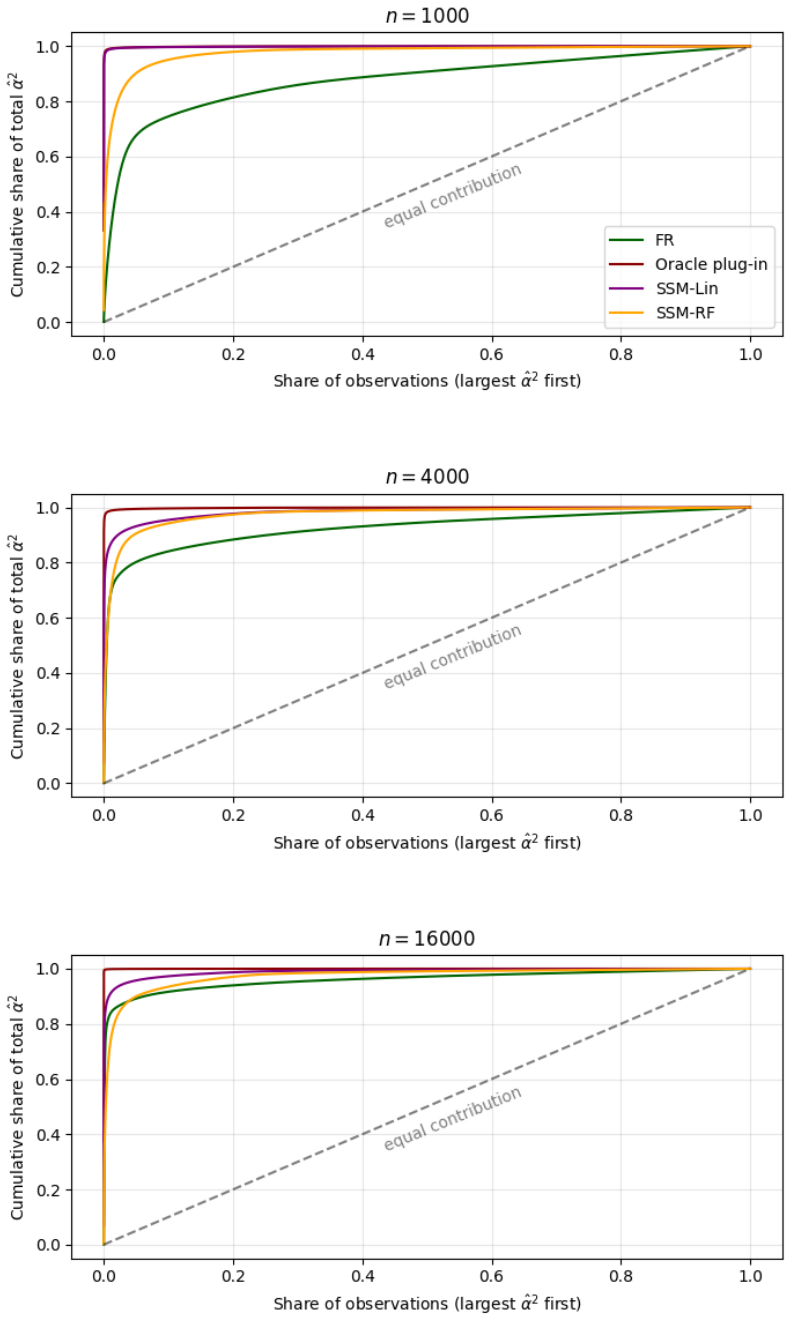
Note: obs_{id} = Observation identifier, Δ = SSM-Lin $-$ Oracle plug-in.

Figure 6: Histograms of $\log(1 + |\hat{\alpha}|)$ | $p = 400$ | $\text{MC}_{\text{Iterations}} = 50$



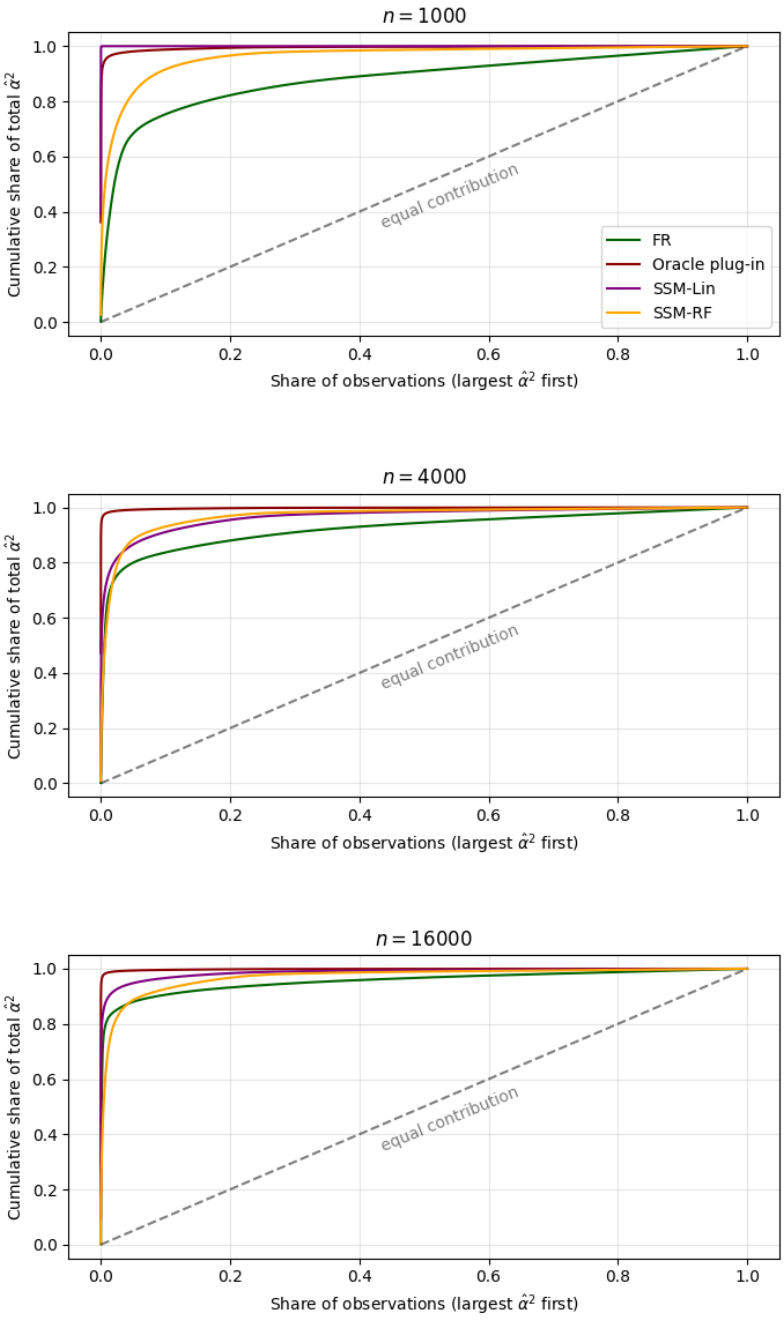
Note: Distributions of the estimated Riesz representer, based on $p = 400$ and 50 Monte Carlo iterations. The figure displays histograms of $\log(1 + |\hat{\alpha}|)$ for different sample sizes and estimation methods under weak overlap.

Figure 7: Concentration curves $\hat{\alpha}^2$ | $p = 100$ | $\text{MC}_{\text{Iterations}} = 200$



Note: Concentration curves of squared Riesz representers under weak overlap for a setting with $p = 100$ covariates, based on 200 Monte Carlo iterations. The curves illustrate the cumulative share of total $\hat{\alpha}^2$ accounted for by the largest estimates (sorted in descending order of $\hat{\alpha}^2$) across sample sizes and estimation methods.

Figure 8: Concentration curves $\hat{\alpha}^2$ | $p = 400$ | $\text{MC}_{\text{Iterations}} = 50$



Note: Concentration curves of squared Riesz representers under weak overlap for a setting with $p = 400$ covariates, based on 50 Monte Carlo iterations. The curves illustrate the cumulative share of total $\hat{\alpha}^2$ accounted for by the largest estimates (sorted in descending order of $\hat{\alpha}^2$) across sample sizes and estimation methods.

Table 7: Tail behavior and second moments of estimated Riesz representers under weak overlap

a) $p = 100$ | $\text{MC}_{\text{Iterations}} = 200$

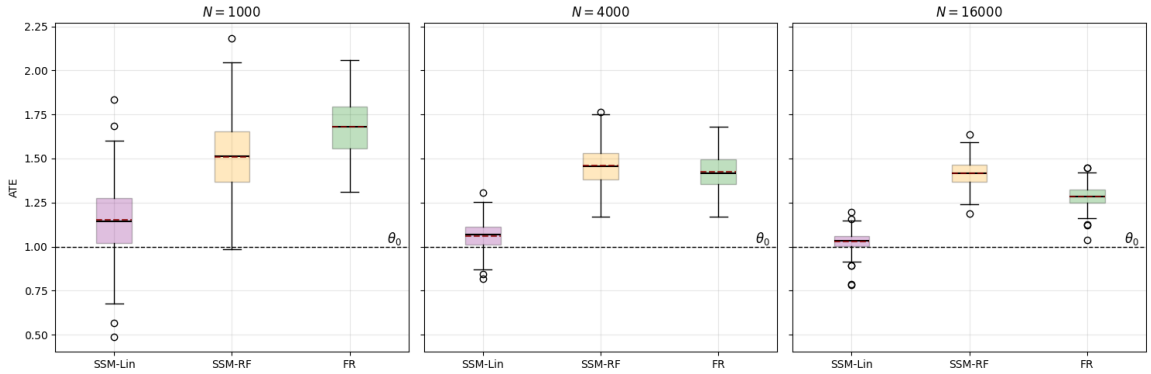
Model	$\mathbf{n}_{\hat{\alpha}}$	$\mathbf{Mean}(\hat{\alpha}^2)$	$\mathbf{Max}(\lvert\hat{\alpha}\rvert)$	$\mathbf{Q.900}(\lvert\hat{\alpha}\rvert)$	$\mathbf{Q.950}(\lvert\hat{\alpha}\rvert)$	$\mathbf{Q.990}(\lvert\hat{\alpha}\rvert)$	$\mathbf{Q.995}(\lvert\hat{\alpha}\rvert)$	$\mathbf{Q.999}(\lvert\hat{\alpha}\rvert)$
				n = 1 000				
FR	112795	6.503	27.496	2.405	4.100	11.611	13.349	17.461
Oracle plug-in	112795	1419.614	7287.958	5.253	8.007	24.913	42.522	160.046
SSM-Lin	112795	1037.886	7175.331	4.567	7.281	24.775	43.126	132.040
SSM-RF	112795	82.043	636.548	6.583	12.279	36.278	51.955	102.361
				n = 4 000				
FR	452775	10.276	65.509	2.413	3.444	12.202	23.621	37.456
Oracle plug-in	452775	1080.990	8657.678	5.290	8.090	25.235	42.153	145.048
SSM-Lin	452775	79.722	1653.624	5.036	7.474	19.886	30.256	81.180
SSM-RF	452775	67.172	375.295	5.691	9.329	37.182	51.828	91.983
				n = 16 000				
FR	1808186	19.197	169.711	2.515	3.793	9.070	15.577	75.188
Oracle plug-in	1808186	9036.138	77986.926	5.281	8.037	24.936	41.932	143.657
SSM-Lin	1808186	134.928	4070.584	5.283	8.050	22.423	34.304	89.720
SSM-RF	1808186	59.255	592.081	5.496	8.331	31.767	48.635	84.152

b) $p = 400$ | $\text{MC}_{\text{Iterations}} = 50$

Model	$\mathbf{n}_{\hat{\alpha}}$	$\mathbf{Mean}(\hat{\alpha}^2)$	$\mathbf{Max}(\lvert\hat{\alpha}\rvert)$	$\mathbf{Q.900}(\lvert\hat{\alpha}\rvert)$	$\mathbf{Q.950}(\lvert\hat{\alpha}\rvert)$	$\mathbf{Q.990}(\lvert\hat{\alpha}\rvert)$	$\mathbf{Q.995}(\lvert\hat{\alpha}\rvert)$	$\mathbf{Q.999}(\lvert\hat{\alpha}\rvert)$
				n = 1 000				
FR	28191	6.659	21.739	2.502	4.062	11.815	13.983	17.191
Oracle plug-in	28191	298.273	1744.652	5.233	7.864	24.322	38.988	126.625
SSM-Lin	28191	141186.889	39584.227	4.530	6.910	25.544	53.141	408.258
SSM-RF	28191	50.308	195.261	6.943	12.242	28.384	39.336	70.464
				n = 4 000				
FR	112747	10.027	45.012	2.337	3.437	13.495	24.323	34.111
Oracle plug-in	112747	693.302	6074.655	5.275	8.137	25.095	40.360	124.610
SSM-Lin	112747	40.526	401.448	5.058	7.442	19.083	27.956	65.812
SSM-RF	112747	56.463	232.134	5.794	9.984	35.447	46.659	76.997
				n = 16 000				
FR	451371	17.047	138.965	2.538	3.677	9.445	16.875	66.925
Oracle plug-in	451371	905.606	8653.149	5.294	8.084	25.202	42.360	141.827
SSM-Lin	451371	108.081	2158.647	5.258	7.875	21.467	33.129	83.650
SSM-RF	451371	53.206	268.204	5.521	8.490	32.643	47.211	76.509

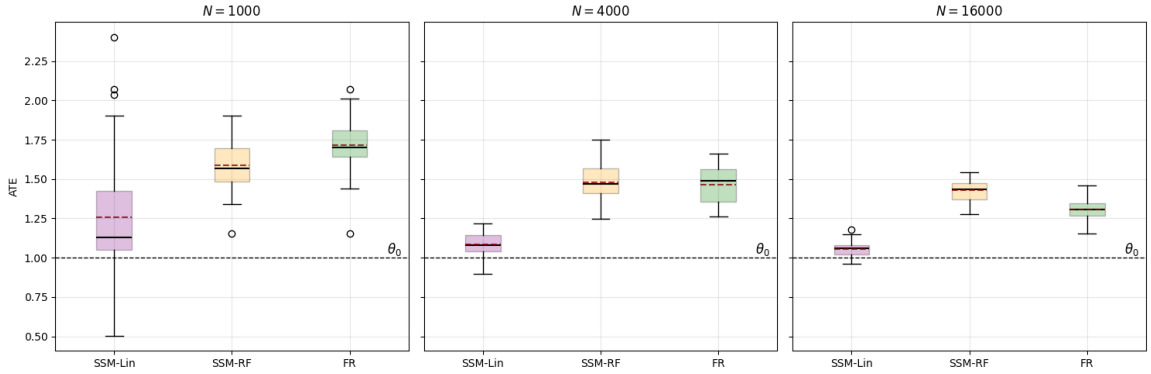
Note: Tail behavior and second moments of estimated Riesz representers under weak overlap for a) a setting with $p = 100$ covariates, based on 200 Monte Carlo iterations, and b) a setting with $p = 400$ covariates, based on 50 Monte Carlo iterations. The table reports the total number of nonzero $\hat{\alpha}$ (selected observations), the mean of $\hat{\alpha}^2$ and the maximum and upper quintiles of $|\hat{\alpha}|$ across estimation methods and sample sizes.

Figure 9: Distribution of $\hat{\theta}$ by estimation approach | $p = 100$ | $\text{MC}_{\text{Iterations}} = 200$



Note: Distribution of estimated average treatment effects under weak overlap for a setting with $p = 100$ covariates, based on 200 Monte Carlo iterations. The figure displays boxplots of $\hat{\theta}$ across estimation methods and sample sizes for the plug-in SSM-DML estimators (linear and random forest) and the Forest Riesz estimator. The dashed red line in each boxplot indicates the mean estimated ATE, while the gray line shows the population parameter value $\theta_0 = 1$.

Figure 10: Distribution of $\hat{\theta}$ by estimation approach | $p = 400$ | $\text{MC}_{\text{Iterations}} = 50$



Note: Distribution of estimated average treatment effects under weak overlap for a setting with $p = 400$ covariates, based on 50 Monte Carlo iterations. The figure displays boxplots of $\hat{\theta}$ across estimation methods and sample sizes for the plug-in SSM-DML estimators (linear and random forest) and the Forest Riesz estimator. The dashed red line in each boxplot indicates the mean estimated ATE, while the gray line shows the population parameter value $\theta_0 = 1$.

Table 8: Hyperparameter and Parameter Setup - Weak Overlap Simulation

Parameters - scikit-learn	**Parameters - doubleML**
RandomForest classes	SSM
n_estimators = 500	n_folds= 3, n_rep= 1
max_depth = 5	score = 'missing-at-random'
min_samples_leaf = 5	normalize_ipw = True
max_features = 0.7	
LassoCV classes	
max_iter = 50 000	
cv = 3	
LogisticRegressionCV classes	
max_iter = 50 000	
penalty = 'l1'	
solver = 'saga'	
cv = 3	

Note: This table reports the hyperparameter setup for the *RandomForestRegressor* and *RandomForestClassifier* classes, as well as the *LassoCV* and *LogisticRegressionCV* classes from *scikit-learn* (Pedregosa et al., 2011), as used within the *DoubleMLSSM* framework of the *doubleML* (Bach et al.) Python package, as well as the *doubleML* parameters. This setup is used to construct the linear and random forest SSM plug-in estimates of the Riesz representer under the weak overlap data-generative process. Parameters not reported are kept at their default values.

Appendix G. Additional Material for the Application

Dependent variable: S (reported wage indicator)

Interaction	**Coef.**	**SE**	**p**
Experience × Female	-0.0242	0.010	0.011**
Experience^2 × Female	0.0007	0.0002	0.001***
Household size × Female	0.0993	0.021	0.000***
Children < 5 × Female	0.0827	0.063	0.190
Master degree × Female	-0.0173	0.058	0.764
Professional degree × Female	0.3203	0.071	0.000***
Doctoral degree × Female	-0.1325	0.113	0.240
Married (absent spouse) × Female	-0.0862	0.174	0.621
Married (present spouse) × Female	-0.1838	0.072	0.011**
Never married × Female	0.1465	0.085	0.083*
Separated × Female	-0.2638	0.217	0.225
Widowed × Female	0.1148	0.229	0.616
Chinese × Female	-0.3191	0.189	0.091*
Other Asian × Female	-0.2551	0.132	0.053*
White × Female	-0.2321	0.091	0.011**
Not well English × Female	0.2016	0.234	0.388
English only × Female	0.4320	0.147	0.003***
English very well × Female	0.4255	0.159	0.007***
English well × Female	0.4326	0.190	0.023**
Hispanic × Female	0.0152	0.111	0.890
Veteran × Female	-0.1159	0.180	0.519
East South Central × Female	0.3031	0.126	0.016**
Middle Atlantic × Female	0.1689	0.086	0.048**
Mountain × Female	0.1045	0.111	0.348
New England × Female	-0.0324	0.104	0.756
Pacific × Female	-0.0938	0.079	0.238
South Atlantic × Female	-0.0907	0.082	0.270
West North Central × Female	0.1876	0.114	0.099*
West South Central × Female	0.0895	0.093	0.335

Table 9: This table presents logit estimates for the probability of reporting wages ($S = 1$). Shown are only the interaction terms between female gender and key socio-economic characteristics. Positive coefficients indicate that the characteristic increases women's reporting probability relative to men, whereas negative coefficients indicate the opposite. The results reveal substantial gender heterogeneity in wage reporting, suggesting that non-random selection into observed wages varies systematically across demographic groups.

Dependent variable: log(wages)

Variable	**Interpretation**
Age	Life-cycle earnings growth
Experience	Linear experience premium
Experience^2	Concavity of returns to experience
College Degree	Returns to education
Married, spouse present	Household stability effect
Professional degree	Very high skill premium
Household size	Family composition
Never married	Labor supply differences
Pacific Division	Regional wage differences
Doctoral degree	Advanced education returns

Table 10: This table reports the covariates most frequently used in the splitting rules of the ForestRiesz regression learner when predicting the outcome Y. Variables with higher split frequency are interpreted as having stronger predictive power for log(wages). The right column provides economic interpretations commonly associated with these predictors.

Group	k	θ_{full}	θ_{-j}	$\Delta\theta$	$\lvert G_{Y,j}\rvert$	$\lvert G_{S,j}\rvert$	$\lvert\rho_j\rvert$
Marital status	5	-0.1276	-0.1331	-0.00553	0.00177	0.00097	1.000
Region	8	-0.1276	-0.1298	-0.00225	0.00110	0.00174	1.000
Race	3	-0.1276	-0.1287	-0.00116	0.00115	0.00264	0.540
Children	1	-0.1276	-0.1278	-0.00027	0.00079	0.00138	0.210
Education	3	-0.1276	-0.1275	0.00007	0.00435	0.01603	0.007
Experience	2	-0.1276	-0.1275	0.00003	0.00003	0.00020	0.284

Table 11: Benchmarking results. Notes: k = number of covariates dropped from group j. θ_{full} = ATE with full covariate set. θ_{-j} = ATE when covariate group j is removed. $\Delta\theta = \theta_{-j} - \theta_{\text{full}}$ measures the sensitivity of the gender wage gap to group j. The sensitivity parameters $G_{Y,j}$, $G_{S,j}$, and ρ_j are defined and explained in Appendix D. Since the estimated sensitivity parameters reported in this table are approximately unbiased for the true shares, some of them can be negative when the true shares are close to zero.

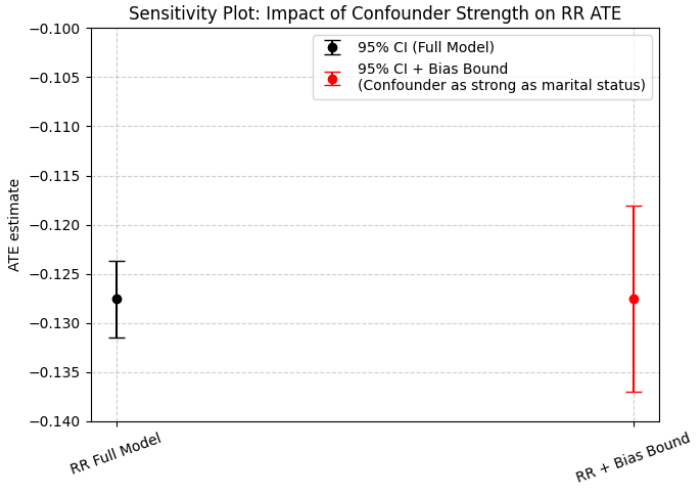


Figure 11: Sensitivity of the estimated gender wage gap (log wages) to potential omitted confounding. The plot compares the Riesz Representer ATE estimate from the full model with the counterfactual confidence interval that would arise if an unobserved confounder were as influential as marital status. While confidence intervals widen under this hypothetical confounder, the estimated ATE remains negative, indicating that the gender wage gap is robust to confounding of realistic magnitude.

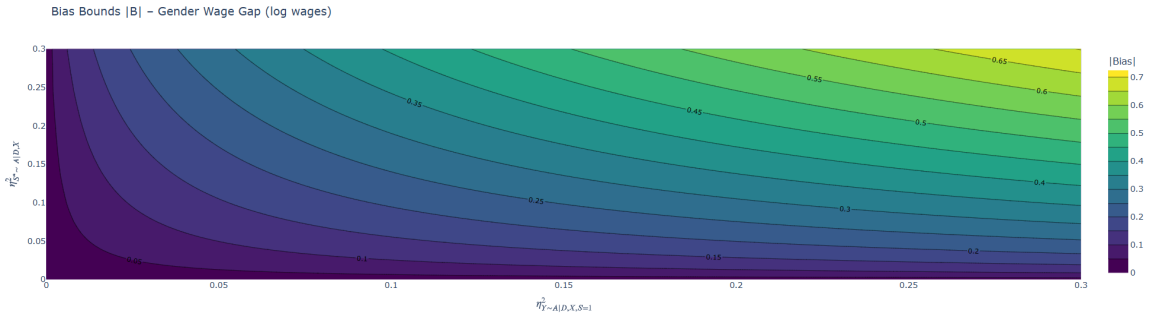


Figure 12: Contour plot of bias bounds as a function of outcome sensitivity $\eta^2_{Y\sim A|D,X,S=1}$ and selection sensitivity $\eta^2_{S^*\sim A|D,X}$. The figure shows how large an omitted confounder must be in terms of explanatory power for both wages and selection into observed wages to overturn the observed gender wage gap. Only confounders with combined sensitivity above the robustness threshold ($RV = 0.063$) could eliminate the estimated effect, implying strong robustness to selection and outcome confounding.