

Robust Second-Stage Double Machine Learning Under Contamination

Daniel K. Baissa
Harvard University
dbaissa@g.harvard.edu

April 28, 2026

Working paper draft. Comments welcome. Please do not circulate without permission.

Abstract

Human data are often complex, nonlinear, and full of interactions, which makes double machine learning an attractive option for social science research. There are two fundamental problems, however. First, even when the underlying data-generating process is clean, the OLS second stage in practical DML is not estimated on oracle Gaussian errors; it is estimated on cross-fitted residuals that combine structural error with first-stage approximation error, fold-level variation, heteroskedasticity, and potentially influential residualized observations. Second, social science data are rarely clean in the first place, and coding uncertainty, measurement error, outliers, leverage points, and distribution shift can further distort the final residual-on-residual regression. This paper asks what happens when the usual OLS second stage of partially linear DML is replaced with a robust estimator. I focus on MM-DML and compare it to OLS-DML, Huber, winsorized, clipped-score, and RANSAC second stages. Across a large simulation benchmark, MM-DML emerges as the strongest general-purpose default. It stays near the clean-data frontier while performing especially well in contamination regimes that are plausible in applied social science. For complex and messy data, researchers should not stop at traditional DML.

1 Introduction

As social scientists studying human behavior, we work with complex data where interactions and nonlinear relationships are the rule rather than the exception, treatment assignment

is rarely clean, and the available controls are often too numerous or too nonlinear for a conventional regression to be credible. Double machine learning (DML) is attractive in exactly these settings because it gives the researcher a disciplined way to use machine learning to adjust for complex controls while still recovering the main effect of interest (Chernozhukov et al., 2018). Yet DML also faces a well-known finite-sample problem. Orthogonality protects the target estimate from first-order nuisance-modeling mistakes, but the final regression is still estimated on feasible, cross-fitted residuals. Those residuals are not oracle errors. They combine the structural disturbance with first-stage approximation error from both nuisance fits, which can create heteroskedasticity, heavy tails, fold-specific variance, and influential residualized observations in the second stage. As a result, the usual intuition that OLS is the natural efficient default is less compelling in finite samples than it first appears.

The same vulnerability becomes sharper when the data themselves are contaminated. Outliers, leverage points, poorly merged records, mismeasured treatments, heavy-tailed outcome noise, weak overlap, and distribution shift can all distort the final estimating equation. In this sense, the finite-sample residualization problem and the contaminated-data problem are not separate concerns. Both operate through the same least durable step in partially linear DML: the second-stage regression of the residualized outcome on the residualized treatment. If that final regression is estimated by least squares, then the familiar weaknesses of least squares remain in place.

Existing work has recognized this fragility in several ways. Some responses keep the DML structure but strengthen the inference layer, using robust variance estimators, repeated sample splitting, or bootstrap procedures. Others alter the score or nuisance-learning strategy to address specific sources of instability, including heavy-tailed outcomes, outliers, weak overlap, and extreme propensity scores (Harada and Fujisawa, 2024; Wang et al., 2024). Still others borrow from robust statistics, using trimming, winsorization, Huber regression, median regression, clipped scores, or related procedures to reduce the influence of unusual observations (Huber, 1964; Hampel, 1974; Yohai, 1987; Huber and Ronchetti, 2009). This paper builds on that literature by targeting the second stage directly. It keeps the standard DML structure, the same cross-fitting logic, and the same causal target, but replaces the least-squares final regression with MM-estimation.

MM-estimation is especially attractive for this role because it was designed to combine resistance to contamination with high efficiency in clean settings (Yohai, 1987; Huber and Ronchetti, 2009). The goal is therefore not only to make DML robust to visibly contaminated data. It is also to make the final stage less sensitive to the irregular residualized observations that can arise from first-stage approximation error itself. This raises two questions. First, can replacing the second-stage OLS regression with an MM-estimator preserve the causal

interpretation of standard DML under the same identifying conditions? Second, if so, do the advantages of MM-estimation carry over to DML when the residualized regression is contaminated, heavy-tailed, or distorted by first-stage approximation error?

Those questions are addressed here in two ways. The first is theoretical, through mathematical results showing when the robust second stage preserves the same causal target. The second is empirical, through a series of simulation experiments. The answer appears to be yes, with qualifications that matter for applied work. In the simulations, MM-DML emerges as the strongest general-purpose robustifier. It ties OLS-DML in RMSE under clean data, which suggests that the efficiency cost of MM is negligible in the feasible DML setting studied here. Under contamination, however, MM-DML improves RMSE under casewise contamination and vertical outliers and is the most frequent winner across the simulated data-generating processes (DGPs). At the same time, the results do not support broad claims of universal dominance. Winsorized and clipped-score procedures are strong specialized competitors under leverage-heavy and treatment-corruption settings, and finite-sample interval behavior remains weaker than the point-estimation story.

The paper then applies MM-DML to a difficult real-world case, examining climate change and migration in the Syrian countryside before the outbreak of the civil war in 2011, using remotely sensed measures of environmental conditions and local activity. The purpose of the empirical illustration is not to provide a definitive test of the climate-conflict hypothesis, but to show what the proposed method recovers in a difficult real-world setting. Using MM-DML, the application finds that areas with poorer soil saw the largest reductions in nighttime lights when the drought began. Establishing a causal link would require much more than this illustration. The onset of the drought would need to be plausibly as-if random, the relevant confounding pathways would need to be controlled for, and the observed decline in nighttime lights would need to be interpretable as migration or related local disruption rather than some unrelated shock. Instead, this application exists to show what an MM-DML workflow can recover in exactly the kinds of difficult observational settings where these substantive debates are usually fought.

The paper proceeds as follows. Section 2 places this research in the DML, robust statistics, and robust causal inference literatures. Section 3 introduces the partially linear setup and develops the identification, orthogonality, and large-sample theory for MM-DML. Section 4 describes the simulation design, including the data-generating processes, contamination mechanisms, first-stage learners, and second-stage estimators. Section 5 presents the simulation results. Section 6 turns to an empirical application in rural Syria. Section 7 offers practical guidance for applied researchers, and Section 8 concludes. Appendix A provides the proof sketch and the theorem-facing summary of the accompanying Lean formalization.

2 From Double Machine Learning to Robust DML

This paper sits at the intersection of three literatures: double/debiased machine learning, robust statistics, and the emerging literature on robustness inside causal procedures. The first explains how machine learning can adjust for complex background relationships while preserving an interpretable causal target. The second explains how estimators can be made less sensitive to outliers, leverage points, heavy tails, and other forms of contamination. The third asks whether those protections can be inserted into causal workflows without changing the estimand.

Double machine learning

The DML framework was developed for settings where a researcher cares about one low-dimensional relationship, but many background covariates also matter and may have complicated functional forms (Chernozhukov et al., 2018). The partially linear model makes this precise. Let Y_i denote the outcome, D_i the treatment, and X_i the observed covariates:

$$Y_i = \theta_0 D_i + g_0(X_i) + U_i, \tag{1}$$

$$D_i = m_0(X_i) + V_i, \tag{2}$$

with $E[U_i | X_i, D_i] = 0$ and $E[V_i | X_i] = 0$. The parameter of interest is the scalar θ_0 . The nuisance functions $g_0(\cdot)$ and $m_0(\cdot)$ absorb potentially high-dimensional or nonlinear covariate relationships that are not themselves the object of interest.

DML estimates these nuisance functions using machine learning, residualizes both the outcome and the treatment on the fitted values, and then estimates θ_0 from the relationship between those residuals. Cross-fitting ensures that each observation is residualized using models trained on held-out data, which prevents overfitting from appearing as first-order bias in the target equation. Neyman orthogonality of the score then ensures that small errors in the nuisance fits enter the estimating equation only at second order (Chernozhukov et al., 2018). The result is a procedure that accommodates flexible first-stage learners while still delivering \sqrt{n} -consistent, asymptotically normal inference on θ_0 .

The standard second stage is OLS on the residualized sample $\{(\tilde{Y}_i, \tilde{D}_i)\}_{i=1}^n$, where any estimator that maps this residual pair to an estimate of θ_0 can be understood as a second-stage DML estimator. OLS-DML is attractive for familiar reasons. It is easy to compute, easy to explain, and efficient under Gaussian conditions. But that same simplicity exposes a vulnerability. When the residualized sample contains gross errors, leverage points, heavy-

tailed errors, or a shifted subgroup, the OLS second stage can be pulled toward observations that are not representative of the main relationship.

There is a further and less obvious difficulty. In the DML pipeline, the second-stage regression does not receive the oracle structural errors as input. The second-stage error for observation i is not U_i alone but the composite residual

$$\varepsilon_i = U_i + (g_0(X_i) - \hat{g}^{(-k(i))}(X_i)) + \theta_0(m_0(X_i) - \hat{m}^{(-k(i))}(X_i)), \quad (3)$$

which combines the structural error with first-stage approximation errors from both nuisance fits. Even under a clean data generating process, these approximation errors are functions of X_i , so $\text{Var}(\varepsilon_i | \tilde{D}_i)$ can vary across observations. The fold structure of cross-fitting adds another source of variance heterogeneity. Observations in different folds are residualized using models trained on different subsets, so estimation quality is not constant across i . The second-stage errors ε_i are therefore heteroskedastic and non-Gaussian in finite samples, even when the underlying data are clean. The Gauss-Markov conditions under which OLS is BLUE do not hold for the input that the second-stage estimator actually receives. This creates the opening for robust second-stage estimators that are designed precisely for influential observations, heavy-tailed residuals, and departures from ideal Gaussian errors.

Robust regression and MM-estimation

Ordinary least squares is sensitive to observations that do not fit the main pattern in the data. This matters especially once DML has transformed the original data into an estimated residualized sample, which violates the Gauss-Markov conditions. Because OLS minimizes squared errors, large residuals receive disproportionate weight, and a small number of unusual observations can substantially distort the fitted line (Huber, 1964; Huber and Ronchetti, 2009). Robust regression methods were developed to reduce this sensitivity while preserving efficiency under clean conditions (Hampel, 1974).

M-estimators address this problem by replacing the squared residual objective with a function that limits the influence of large residuals. The estimator iteratively reweights observations, assigning lower weight to those with unusually large residuals relative to the current fit, until convergence. Huber regression is the most familiar case. It behaves like OLS for small residuals but limits influence once residuals exceed a threshold (Huber, 1964). Redescending score functions go further, allowing extreme residuals to receive near-zero weight (Beaton and Tukey, 1974).

High-breakdown estimators ask a stronger question. How much of the sample can

be contaminated before the estimator is driven arbitrarily far from the truth? OLS has a breakdown point near zero. S-estimators and related procedures achieve breakdown points near 50% (Rousseeuw, 1984; Rousseeuw and Yohai, 1984), but at a cost in efficiency under clean conditions.

MM-estimation combines both goals (Yohai, 1987). It begins with a high-breakdown S-estimator to obtain a resistant starting point, then applies an efficient M-estimation step around that starting point. The result achieves a high breakdown point and near-OLS efficiency simultaneously. The standard 95% asymptotic relative efficiency result for MM is derived relative to OLS under symmetric Gaussian errors. In the DML setting, however, the second-stage errors are estimated residual objects rather than ideal Gaussian disturbances. The relevant comparison is therefore not whether MM sacrifices a small amount of efficiency under ideal errors, but whether that sacrifice persists after cross-fitting and first-stage approximation have already made the residualized sample heteroskedastic and non-Gaussian.

This helps explain a finding from the simulations reported below. MM and OLS achieve nearly identical RMSE under no contamination (Table 5). OLS holds no visible efficiency advantage once the DML pipeline has already altered the second-stage error distribution, while MM retains its protection against contaminated residualized observations. The practical implication is that replacing OLS with MM in the second stage costs little or nothing in the clean case while providing insurance in the contaminated cases that follow.

Other robust procedures offer complementary protections. Winsorization clips extreme values of the residualized outcome or treatment before fitting (Tukey and McLaughlin, 1963; Dixon and Yuen, 1974). Clipped-score procedures instead truncate the empirical score contributions directly, targeting observations whose residualized outcome and treatment interact in especially damaging ways. RANSAC treats the second stage as an inlier-selection problem, repeatedly fitting on subsets and choosing the fit supported by the largest consistent inlier set (Andersen, 2008). These procedures are less central to the theory developed below, but they serve as practical benchmarks for applied researchers choosing among second-stage options.

Robustness inside causal procedures

The growing literature on robustness inside causal procedures asks a narrower question than classical robust statistics. It asks whether robustness can be inserted into estimators that applied researchers already use for causal inference without changing the estimand or obscuring interpretation (Harada and Fujisawa, 2024; Wang et al., 2024). This distinction matters because robust standard errors and robust estimation solve different problems.

Heteroskedasticity-consistent covariance estimators protect inference against misspecified variances but do not repair a contamination-sensitive estimating equation (White, 1980; Freedman, 2006; King and Roberts, 2015). In the present setting, the question is concrete. Can the OLS second stage of partially linear DML be replaced by an MM-style robust score without redefining the causal target? The next section shows that the answer is yes, under explicit and verifiable conditions.

3 Identification and Orthogonality Under Robustification

This section answers the question raised at the end of the previous section: can the second stage of partially linear DML be robustified without losing the causal meaning of the original estimator? The answer is conditional and deliberately narrow. Robustification does not itself identify a causal effect. Identification still comes from the partially linear causal structure. The role of the theory is to show what must remain true when the usual least-squares second stage is replaced by an MM-style robust score.

The mathematics below should therefore be read as a synthesis rather than a new theory built from scratch. It takes the standard partially linear DML score from Chernozhukov et al. (2018) and replaces the least-squares residual score with the robust score functions used in M- and MM-estimation, especially the MM framework developed by Yohai (1987). The goal is not to rederive all of DML or all of robust statistics. It is to state the bridge needed for this paper: under explicit centering, local-identification, orthogonality, and regularity conditions, the robustified second stage continues to target the same partially linear parameter. A fuller proof sketch, including the corresponding formal theorem stack, is provided in Appendix A.

The starting point is the same partially linear model used in standard DML:

$$Y_i = \theta_0 D_i + g_0(X_i) + U_i, \tag{4}$$

$$D_i = m_0(X_i) + V_i, \tag{5}$$

with $E[U_i | X_i, D_i] = 0$ and $E[V_i | X_i] = 0$. The parameter of interest is θ_0 . The nuisance functions $g_0(\cdot)$ and $m_0(\cdot)$ absorb the potentially complicated relationship between the covariates, the outcome, and the treatment. Standard DML estimates nuisance functions flexibly, residualizes the outcome and treatment, and then estimates θ_0 from the resulting residualized sample.

For the orthogonal partially linear score, it is useful to distinguish the structural outcome component $g_0(X)$ from the outcome regression used for residualization. Define

$$\ell_0(X) := E[Y | X]. \quad (6)$$

Under the partially linear model,

$$\ell_0(X) = \theta_0 m_0(X) + g_0(X). \quad (7)$$

This distinction matters because the DML score residualizes Y using $\ell_0(X)$, not the structural component $g_0(X)$ alone.

The robustified version keeps this structure but changes the final score. Write the population robust score as

$$\Psi(\theta, \ell, m, s) = E \left[(D - m(X)) \psi \left(\frac{Y - \ell(X) - \theta(D - m(X))}{s} \right) \right]. \quad (8)$$

When $\psi(u) = u$, this is the usual least-squares residual score. For MM-DML, ψ is instead a robust score associated with the M-estimation refinement of the MM procedure. The target parameter remains θ_0 ; what changes is how second-stage residual contributions enter the estimating equation.

At the truth, define the residualized treatment and orthogonal structural residual by

$$V = D - m_0(X), \quad (9)$$

$$U = Y - \ell_0(X) - \theta_0 V. \quad (10)$$

The score at the truth is therefore

$$\Psi(\theta_0, \ell_0, m_0, s_0) = E \left[V \psi \left(\frac{U}{s_0} \right) \right]. \quad (11)$$

This expression highlights the key issue. For the usual OLS score, centering follows from the familiar residual moment. For a general robust score, centering is not automatic. It must either be derived from additional distributional structure or imposed as an explicit condition. The theory below takes the latter route. This makes the claim transparent: MM-DML preserves the standard DML target under conditions that ensure the robust score is centered at that same target.

The proof logic has four steps. First, the robust population score must equal zero at θ_0 . Second, movement in θ must move the population score locally, so that θ_0 is identified

rather than merely compatible with the score. Third, the score must remain orthogonal to first-order changes in the nuisance functions, so that cross-fitted machine-learning error enters only at higher order. Fourth, the empirical cross-fitted score must admit a local expansion that yields consistency, asymptotic linearity, and the usual inference layer under additional variance-control conditions.

Assumption 3.1 (Partially linear causal structure). The data satisfy the partially linear model above, with $E[U_i | X_i, D_i] = 0$, $E[V_i | X_i] = 0$, and non-degenerate residualized treatment variation.

Assumption 1 is the structural backbone of the paper. It is what gives θ_0 causal meaning in the first place. Nothing about MM-estimation or robustification creates that interpretation. The paper keeps the usual partially linear target and studies whether the final stage can estimate it more reliably when the residualized sample is contaminated.

Assumption 3.2 (Robust score centering). At the true nuisance functions and scale, the robust population score is centered at the partially linear target:

$$E\left[V\psi\left(\frac{U}{s_0}\right)\right] = 0. \quad (12)$$

Assumption 2 is the condition that prevents robustification from silently changing the estimand. It says that, after replacing the least-squares residual score with a robust score, the population moment still vanishes at θ_0 . This is the key distinction between making a regression robust and preserving a causal target. A robust score can reduce the influence of contaminated residuals, but it remains a DML estimator of the original treatment effect only if its population moment is centered at that effect.

A simple sufficient condition is the classical symmetric-error case. If, conditional on the residualized treatment and covariates, U/s_0 has a distribution symmetric around zero and the robust score ψ is odd, then $E[\psi(U/s_0) | V, X] = 0$, which implies

$$E\left[V\psi\left(\frac{U}{s_0}\right)\right] = 0. \quad (13)$$

This is the standard centering logic behind many robust M-estimators with odd score functions (Huber and Ronchetti, 2009).

Assumption 3.3 (Local identification). The population score is differentiable in the target direction at the truth, with Jacobian

$$J := \partial_\theta \Psi(\theta, \ell_0, m_0, s_0)\Big|_{\theta=\theta_0}, \quad (14)$$

and $J \neq 0$.

Assumption 3 is the local-identification condition. Intuitively, once the residualized treatment still contains variation after conditioning on X , movement in θ should change the population score locally. The nonzero Jacobian condition ensures that the score isolates θ_0 in the target direction rather than being flat or uninformative near the truth (Newey and McFadden, 1994; van der Vaart, 1998).

Assumption 3.4 (Neyman orthogonality). For admissible perturbations h and q of the nuisance functions ℓ_0 and m_0 , the first-order nuisance derivatives vanish at the truth:

$$\partial_\ell \Psi(\theta_0, \ell_0, m_0, s_0)[h] = 0, \tag{15}$$

$$\partial_m \Psi(\theta_0, \ell_0, m_0, s_0)[q] = 0. \tag{16}$$

Assumption 4 is the DML part of the argument. The point is not that nuisance estimation is perfect. The point is that the score is constructed so that first-order nuisance errors do not drive the treatment effect estimate. Without this orthogonality property, there would be little reason to expect a robust second stage to retain the same large-sample logic as standard DML (Chernozhukov et al., 2018; Newey and McFadden, 1994).

At the population level, Assumption 4 states the orthogonality property itself. At the sample level, this property must be paired with the usual DML rate conditions on the first-stage nuisance estimates. Those sample-level requirements are stated formally below in Assumption 3.7.

Assumption 3.5 (Robust score and scale regularity). The score function ψ is regular enough near the truth to support the differentiability, orthogonality, and local expansion arguments above, and the scale estimator is consistent for s_0 .

Assumption 5 connects robust statistics back to semiparametric identification. A bounded or redescending score is useful only if it still behaves stably near the truth. The score may reduce the influence of extreme residuals, but it must remain sufficiently regular for the population and sample expansions to be meaningful (Huber and Ronchetti, 2009; van der Vaart, 1998).

For example, the Tukey biweight score used in many MM-estimation procedures is bounded, redescending, and differentiable except possibly at cutoff points that can be handled under standard smoothness or negligible-boundary-mass conditions. Thus the regularity requirement should be read as the familiar robust-statistics condition that the chosen score be stable enough to support local Taylor expansion while still bounding the influence of extreme residuals (Yohai, 1987; Huber and Ronchetti, 2009).

Theorem 3.6 (Population targeting and orthogonality of MM-DML). *Under Assumptions 1–5, the MM-DML population score satisfies*

$$\Psi(\theta_0, \ell_0, m_0, s_0) = 0, \tag{17}$$

locally identifies θ_0 through a nonzero target-direction Jacobian J , and is Neyman-orthogonal to first-order perturbations in the nuisance functions ℓ_0 and m_0 .

Theorem 3.6 is the point at which the paper’s claim becomes precise. It does not say that MM-DML creates causal identification. It says that, under the same partially linear causal structure used by standard DML and under an explicit robust-score centering condition, robustifying the second stage does not define a new estimand. This separates two questions that are often conflated in practice. One question is whether the method is robust. The other is whether it still estimates the same causal object. The theorem answers the second question first.

The theorem also clarifies why the robust score is not merely a descriptive regression device. At the population level, the robustified score has the same target-zero property as the usual DML score, has enough curvature in the target direction to identify θ_0 , and preserves the nuisance-insensitivity that makes DML useful with flexible first-stage learners. Those are the formal ingredients needed for the robust second stage to remain part of the DML framework rather than becoming a different estimator with a different target.

The sample estimator is defined from the cross-fitted residuals

$$\tilde{Y}_i = Y_i - \hat{\ell}^{(-k(i))}(X_i), \tag{18}$$

$$\tilde{D}_i = D_i - \hat{m}^{(-k(i))}(X_i), \tag{19}$$

through the empirical robust score

$$\hat{\Psi}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \tilde{D}_i \psi \left(\frac{\tilde{Y}_i - \theta \tilde{D}_i}{\hat{s}} \right). \tag{20}$$

The MM-DML estimator is any local approximate root satisfying

$$|\hat{\Psi}_n(\hat{\theta}_n)| \leq \text{tol}_n. \tag{21}$$

This definition matters because it mirrors the computational object used in the simulations. The nuisance fits are estimated with cross-fitting, the residualized sample is held fixed, and the second-stage robust score determines the final estimate.

The primitive bundles used in the appendix, including the `PrimitiveScoreApproxAssumptions` and `PrimitiveLinearizationAssumptions`, encode these same DML requirements at the level of the formal proof skeleton: score approximation, nuisance-product negligibility, local differentiability, and variance control. They are not additional identifying assumptions separate from the estimator; they formalize the standard rate and expansion conditions needed to move from the population robust score to the cross-fitted sample estimator.

Assumption 3.7 (Score approximation and nuisance-product rates). The empirical cross-fitted score converges uniformly enough to its population counterpart near θ_0 , the approximate-root tolerance satisfies $\text{tol}_n = o_p(1)$, and the population score is locally identified at θ_0 . In addition, the first-stage nuisance estimators satisfy the standard DML product-rate condition, so that products of nuisance errors are asymptotically negligible. In the outcome-regression form used by the orthogonal score, this requires terms of the form

$$\|\hat{\ell} - \ell_0\| \cdot \|\hat{m} - m_0\| = o_p(n^{-1/2}). \quad (22)$$

Equivalently, using the structural outcome nuisance g_0 from the partially linear model, the corresponding condition can be written as

$$\|\hat{g} - g_0\| \cdot \|\hat{m} - m_0\| = o_p(n^{-1/2}), \quad (23)$$

up to the usual relation $\ell_0(X) = \theta_0 m_0(X) + g_0(X)$. Verification of these rates for particular learners follows the standard DML arguments for cross-fitted nuisance estimation under appropriate complexity and regularity conditions; it is not a new learner-specific requirement introduced by MM-DML ([Chernozhukov et al., 2018](#)).

Proposition 3.8 (Consistency of cross-fitted MM-DML). *Under Assumptions 1–6, any local approximate root $\hat{\theta}_n$ of the cross-fitted MM-DML score satisfies*

$$\hat{\theta}_n \xrightarrow{p} \theta_0. \quad (24)$$

The consistency statement is the sample analogue of the population targeting result. Once the robust population score identifies θ_0 , and once the cross-fitted empirical score approximates that population score well enough, an approximate empirical root must concentrate near the same target. The role of cross-fitting is the same as in standard DML: it helps prevent first-stage overfitting from appearing as first-order movement in the target equation.

Assumption 3.9 (Linearization, variance control, and studentization). The empirical cross-

fitted robust score admits a local expansion around θ_0 with an $o_p(n^{-1/2})$ remainder, and the studentizing factor consistently estimates the asymptotic variance of the leading term.

Assumption 7 is stronger than the earlier identification and consistency conditions because inference is harder than point estimation. To show that a robust second stage converges to the correct target, one mainly needs score centering, local identification, and nuisance control. To justify Wald-style intervals, one also needs a stable local expansion and variance control (Newey and McFadden, 1994; van der Vaart, 1998). The simulations later in the paper help explain why that difference matters in finite samples.

In implementation, this studentization condition need not rely exclusively on a closed-form analytic variance estimator. A pairs bootstrap applied to the cross-fitted residualized sample provides a practical route for approximating the sampling distribution of the robust second-stage estimator while preserving the fitted first-stage structure. When finite-sample contamination makes analytic Wald intervals unreliable, a pairs bootstrap applied to the cross-fitted residualized sample is the preferred practical route, as it approximates the sampling distribution of the robust second-stage estimator while preserving the fitted first-stage structure.

Theorem 3.10 (Asymptotic linearity of cross-fitted MM-DML). *Under Assumptions 1–7, any local approximate root $\hat{\theta}_n$ of the cross-fitted MM-DML score satisfies*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = J^{-1}L_n + o_p(1), \quad (25)$$

where J is the population Jacobian and L_n is the leading score-at-truth term.

Theorem 3.10 says that the large-sample logic of DML survives the robustification step under the stated conditions. Once orthogonality removes first-order nuisance effects, the estimator is driven by a leading term built from the score at the truth. What changes relative to OLS is the behavior of that score contribution under contamination. With OLS, extreme residualized observations can have unbounded influence. With MM-DML, the score is designed to reduce that influence.

Corollary 3.11 (Asymptotically valid Wald inference for MM-DML). *Under the conditions of Theorem 3.10 and consistent studentization, the studentized MM-DML estimator is asymptotically normal and supports asymptotically valid Wald-style confidence intervals and hypothesis tests for θ_0 .*

The corollary should be read carefully. It is a large-sample statement, not a promise of uniformly well-behaved finite-sample inference under every contamination mechanism. That

distinction matters in this paper. The benchmark shows a stronger point-estimation story than interval story, especially in the hardest treatment-contamination settings. That is not a contradiction of the theory. It is a reminder that the asymptotic logic is cleaner than the finite-sample environment.

It is also important to be precise about scope. The main theorems are about MM-DML because MM is the robust score-based estimator around which the large-sample argument is most naturally built. Winsorized, clipped-score, and RANSAC procedures are still important benchmark competitors, but they are included primarily as practical comparators rather than as the main objects of the semiparametric theory. The appendix records the more complete proof sketch and the formal theorem stack, including the distinction between asymptotic linearity and the proxy/event-lifting arguments used for the inference layer.

4 Simulation Design

This section lays out the design choices for the simulation used to test the performance of MM-DML against the second-stage alternatives considered in the paper. It begins with a high-level summary of the full design and then turns to the individual pieces in more detail. It first describes the data-generating processes, then the contamination mechanisms, then the matched first-stage learners and second-stage estimators, and finally the overall simulation workflow.

Table 1: Simulation design

| Component | Design choice |
|---------------------------|---|
| Target parameter | $\theta_0 = 1$ |
| Sample sizes | $n \in \{300, 600\}$ |
| Covariate dimension | $p = 20$ |
| Data-generating processes | Sparse linear; nonlinear interactions; regime-shift nonlinear |
| First-stage learners | Lasso; elastic net; random forest; histogram gradient boosting |
| Contamination mechanisms | None; vertical outliers; leverage points; treatment contamination; cellwise contamination; casewise contamination; domain shift |
| Contamination fractions | $\{0, 0.01, 0.05, 0.10, 0.25\}$ |
| Contamination magnitudes | $\{5, 10\}$ |
| Second-stage estimators | OLS; Huber; MM; winsorized; clipped-score; RANSAC |
| Monte Carlo replications | 200 per design cell |

Note: Within each grouped scenario, the residualized sample is computed once and the same cross-fitted residual pair is passed to every second-stage estimator. This holds the first stage fixed across competitors and isolates second-stage robustness.

Table 1 gives the simulation design at a high level before the section turns to the individual pieces in more detail. The main design choice is that, within each grouped scenario, the residualized sample is computed once and then passed unchanged to each second-stage estimator. That keeps the comparison focused on the second stage itself rather than on incidental differences in nuisance fitting. The rest of the section then moves from that overview to the lower-level design choices: the DGPs, the contamination mechanisms, the matched first-stage learners, the second-stage estimators, and the simulation workflow.

The full simulation slate contains 588 grouped scenarios, 3,528 scenario-method cells, and 705,600 method-level results. That scale matters because it means the main comparisons are not riding on a handful of replications or on one especially favorable design cell. At the same time, the design is still narrow enough to stay interpretable. The point is not to test every possible failure mode or every possible learner combination. It is to study a limited set of questions carefully and under conditions that are broad enough to be useful. With that high-level structure in view, the rest of the section turns to the design pieces one at a time, beginning with the data-generating processes.

4.1 Data-Generating Processes

4.1.1 Common covariate design

All three DGPs begin from the same covariate distribution,

$$X \sim \mathcal{N}(0, \Sigma), \quad \Sigma_{jk} = 0.3^{|j-k|},$$

so that the design includes correlation across covariates without becoming unnecessarily complicated. That choice gives the simulations a common starting point that is simple enough to interpret but still more realistic than treating the covariates as independent. Treatment and outcome are then generated from the partially linear model

$$D = m_0(X) + \eta, \quad \eta \sim N(0, 1), \tag{26}$$

$$Y = \theta_0 D + g_0(X) + \varepsilon, \quad \varepsilon \sim N(0, 1), \tag{27}$$

with $\theta_0 = 1$. That keeps the causal target fixed across the simulations and keeps the comparison focused on the same treatment effect. What changes from one design to the next is the nuisance structure through the form of $g_0(X)$ and $m_0(X)$.

4.1.2 DGP 1: sparse linear

The first DGP is the simplest of the three and serves as the clean baseline. It is called sparse because only a few covariates enter the nuisance functions at all. In this design, X_1 , X_2 , and X_3 help predict treatment assignment and the outcome, while the remaining covariates play no direct role. It is called linear because those covariates enter through simple weighted sums rather than through interactions, thresholds, or other nonlinear transformations. Concretely,

$$g_0(X) = 1.0X_1 - 0.8X_2 + 0.6X_3, \tag{28}$$

$$m_0(X) = 0.8X_1 + 0.4X_2 - 0.5X_3. \tag{29}$$

This means that the outcome nuisance is increasing in X_1 and X_3 but decreasing in X_2 , while the treatment nuisance is increasing in X_1 and X_2 but decreasing in X_3 . The coefficients also make X_1 the strongest common driver of both treatment and outcome, which creates confounding in exactly the way the DML setup is meant to address.

This design is intentionally favorable to sparse linear first-stage learners such as lasso and elastic net. That is useful because it gives the benchmark a setting in which the nuisance problem is well specified, or at least close to it. In that setting, the comparison is not about

whether one method can rescue a badly misspecified first stage. Instead, it provides a clean baseline for asking a narrower question: when the nuisance stage is behaving as it should, do robust estimators in the second-stage pay a meaningful price relative to OLS, or can they remain competitive even before contamination is introduced?

4.1.3 DGP 2: nonlinear interactions

The second DGP moves away from the clean sparse-linear case and introduces a more complicated set of relationships between the covariates, treatment assignment, and the outcome in ways researchers will recognize more readily from applied work. Here, treatment assignment and the outcome are no longer driven by simple weighted sums. Instead, they depend on nonlinear transformations of individual covariates and on interactions between covariates. Concretely,

$$g_0(X) = 1.5 \sin(X_1) + 0.8X_2X_3 - 0.6X_4^2 + 0.4 \exp(\text{clip}(X_5/2, -2, 2)), \quad (30)$$

$$m_0(X) = 0.8 \cos(X_1) + 0.5X_2X_4 - 0.4X_3^2 + 0.3 \tanh(X_5). \quad (31)$$

The outcome equation is built from four different pieces. First, X_1 enters through $\sin(X_1)$, so its effect is nonlinear and changes over its range. Second, X_2 and X_3 enter together through an interaction term, so their contribution depends on their joint values rather than on either variable alone. Third, X_4 enters quadratically through X_4^2 , which introduces curvature. Fourth, X_5 enters through a clipped exponential term, which allows a nonlinear effect without letting it grow without bound.

The treatment equation is built in a similar spirit, but it is not the same function. It uses $\cos(X_1)$ rather than $\sin(X_1)$, includes an interaction between X_2 and X_4 rather than between X_2 and X_3 , adds a negative quadratic term in X_3^2 , and lets X_5 enter through $\tanh(X_5)$.

In plain language, this DGP is designed so that the same covariates matter for treatment and outcome, but they matter in more complicated ways than in the sparse-linear case. Some effects are nonlinear, some depend on pairs of variables moving together, and some are explicitly curved rather than monotone. That creates confounding that is still systematic, but no longer well described by a sparse linear regression. A learner that only sees straight-line additive structure will therefore be at a disadvantage.

That is the point of including this design. It creates a setting in which nonlinear first-stage learners such as random forests and histogram gradient boosting are on more natural

ground, while linear learners are no longer privileged by construction. In turn, this lets the benchmark ask whether robust second-stage procedures still help once the nuisance problem itself is more realistic and less forgiving than the clean sparse-linear baseline. In other words, DGP 2 moves the benchmark closer to the kinds of nonlinear control relationships that often motivate DML in the first place.

4.1.4 DGP 3: regime-shift nonlinear

The third DGP keeps the nonlinear character of DGP 2, but the relationship between the covariates, treatment assignment, and the way the covariates affect treatment assignment and the outcome now differs across the sample. In this design, the functional form depends on the sign of X_1 . When $X_1 \leq 0$, the model uses the same equations as DGP 2. When $X_1 > 0$, it switches to

$$g_0(X) = 1.2 \cos(X_1) - 0.7X_2X_3 + 0.5|X_4| + 0.4 \tanh(1.5X_5), \quad (32)$$

$$m_0(X) = 0.7 \sin(X_1) + 0.6X_2X_5 - 0.3X_3^2 + 0.4 \exp(\text{clip}(X_4/3, -2, 2)). \quad (33)$$

So this is not just a nonlinear design. It is a design in which the rules themselves shift depending on where an observation falls. Two observations that look similar in most covariates can therefore be generated by different underlying relationships if they fall on different sides of the $X_1 = 0$ cutoff.

The outcome equation in the $X_1 > 0$ region combines four different features. It uses $\cos(X_1)$ rather than $\sin(X_1)$, reverses the sign on the interaction between X_2 and X_3 , replaces the quadratic term in X_4 with the absolute-value term $|X_4|$, and lets X_5 enter through $\tanh(1.5X_5)$. The treatment equation also changes. It uses $\sin(X_1)$ rather than $\cos(X_1)$, replaces the earlier interaction with X_2X_5 , keeps a negative quadratic term in X_3^2 , and lets X_4 enter through a clipped exponential term.

In plain language, this DGP is meant to capture a setting where the same variables do not affect treatment and outcome in exactly the same way everywhere in the sample. Instead, part of the sample follows one set of relationships and another part follows a different one. That kind of regime dependence is common in applied work. Political behavior, economic activity, or administrative processes often do not follow one stable pattern across all cases.

That is why this design matters for the benchmark. It creates a setting with structured heterogeneity rather than isolated outliers. The challenge is not simply that some observations are extreme. It is that one part of the sample is generated by a different under-

lying pattern than another. That makes DGP 3 especially useful for thinking about domain shift and for asking whether robust second-stage procedures still help when the data are not just nonlinear, but locally unstable across the covariate space.

4.2 Contamination Mechanisms

The benchmark studies seven contamination mechanisms, each meant to capture a different way messy data can distort the residualized regression after the nuisance stage has been estimated. Some alter the outcome directly, some create leverage, some corrupt treatment measurement, and some shift the structure of the data more broadly.

4.2.1 No contamination

The clean benchmark leaves $(X,D,Y)(X,D,Y)(X,D,Y)$ unchanged. This provides the baseline against which the contaminated designs are judged. If a robust second-stage estimator is going to be practically useful, it should remain competitive here as well, rather than improving performance only by sacrificing too much in clean data.

4.2.2 Vertical outliers

In the vertical-outlier design, selected observations receive a shock to the outcome while the covariates and treatment are otherwise left in place:

$$Y_i \leftarrow Y_i + \text{magnitude} \cdot \text{sign}(Z_i), \quad Z_i \sim N(0, 1).$$

This creates observations whose outcomes are unusually high or low relative to what the rest of the data would suggest. The point is to study the classic case in which the problem lies mainly in the response rather than in the design matrix. This matters because it is the setting robust regression methods are often introduced to handle, so it provides a natural first test of whether those same protections still help once the second stage is embedded inside a DML workflow.

4.2.3 Leverage points

In the leverage-point design, selected observations are pushed away from the rest of the sample in the covariates, and they also receive smaller shifts in treatment and outcome:

$$X_i \leftarrow X_i + \text{magnitude} \cdot \text{sign}(Z_i^{(x)}), \quad (34)$$

$$D_i \leftarrow D_i + 0.25 \cdot \text{magnitude} \cdot \text{sign}(Z_i^{(d)}), \quad (35)$$

$$Y_i \leftarrow Y_i + 0.25 \cdot \text{magnitude} \cdot \text{sign}(Z_i^{(y)}). \quad (36)$$

In this benchmark, leverage points are selected adversarially, meaning that the contaminated observations are those with the largest deviations of D from its median.

In plain language, this creates observations that sit in unusual parts of the design space rather than simply having unusual outcomes. Those observations can matter a great deal in a regression because they have the potential to pull the fitted line toward themselves. That is what makes leverage different from a vertical outlier. The problem is not just that an observation looks strange in the response. It is that its covariate values give it unusual geometric influence.

This case matters because many real data problems look more like leverage than like pure outcome outliers. A bad merge, miscoded covariate, extreme scale difference, or unusual subgroup can create observations that are not obviously wrong in the outcome alone but still exert disproportionate influence on the fitted regression. Including this design therefore makes it possible to see whether second-stage methods that clip or downweight influential points are especially helpful when the distortion comes through the geometry of the regression rather than through the outcome by itself.

4.2.4 Treatment contamination

In the treatment-contamination design, selected observations receive a shock to the treatment while the rest of the data-generating process is left in place:

$$D_i \leftarrow D_i + \text{magnitude} \cdot \text{sign}(Z_i), \quad Z_i \sim N(0, 1).$$

In plain language, this creates observations in which the treatment is measured incorrectly or is otherwise pushed away from the value implied by the underlying covariates. That makes this case different from vertical outliers and leverage points. Here the problem is not mainly that the outcome looks unusual or that an observation sits in an unusual part of the design space. The problem is that the regressor of interest itself has been corrupted.

This case matters because it marks an important boundary of what second-stage robustness can fix. A robust second-stage estimator can downweight influential observations,

but it cannot fully recover information that has been directly damaged in the treatment variable itself. In applied work, that makes this contamination mechanism especially relevant for settings with miscoded treatments, reporting errors, timing mismatches, or other forms of measurement error in the variable whose effect the researcher is trying to estimate.

4.2.5 Cellwise contamination

In the cellwise-contamination design, each selected observation has one randomly chosen covariate entry perturbed:

$$X_{ij_i} \leftarrow X_{ij_i} + \text{magnitude} \cdot \text{sign}(Z_i),$$

where j_i is a randomly chosen coordinate.

In plain language, this means that the whole row is not corrupted. Instead, one value inside the row is wrong. That is a useful contrast with casewise contamination, where the entire observation is damaged. Here the observation is mostly intact, but one covariate has been miscoded, shifted, or recorded incorrectly.

This case matters because many real data problems are cellwise rather than casewise. A single survey answer can be miscoded, one administrative field can be entered incorrectly, or one merged variable can be off even when the rest of the record is fine. That kind of contamination is often mild enough that it does not make an observation look obviously broken, but it can still distort the fitted regression, especially once the nuisance stage has been estimated and the comparison turns on the geometry of the residualized sample.

Including this design therefore makes it possible to study a more ordinary form of dirty data. The point is not to create spectacular outliers, but to see how methods behave when observations are mostly usable yet contain localized errors that can still affect treatment assignment and the outcome indirectly through the covariates.

4.2.6 Casewise contamination

In the casewise-contamination design, selected observations are corrupted at the row level:

$$X_i \leftarrow X_i + \text{magnitude} \cdot Z_i^{(x)}, \tag{37}$$

$$D_i \leftarrow D_i + \text{magnitude} \cdot Z_i^{(d)}, \tag{38}$$

$$Y_i \leftarrow Y_i + \text{magnitude} \cdot Z_i^{(y)}, \tag{39}$$

In plain language, this means the whole observation is damaged rather than just one field inside it. The covariates, treatment, and outcome are all shifted at once. That makes this the clearest simulation analogue of a bad record. A row might be badly merged, assigned to the wrong unit, recorded with multiple errors, or otherwise fail in a way that affects more than one variable at the same time.

This case matters because many of the ugliest data problems in applied work are casewise rather than cellwise. When an observation is corrupted across several variables at once, ordinary least squares can be especially fragile because the same bad row can distort both the geometry of the regression and the apparent relationship between treatment and outcome. Including this design therefore makes it possible to test whether robust second-stage estimators are especially helpful when the problem is not a single miscoded value, but a whole observation that should not be trusted in its recorded form.

4.2.7 Domain shift

In the domain-shift design, contamination is introduced as a structured drift rather than as a small collection of isolated bad points:

$$X_{i1} \leftarrow X_{i1} + 0.5 \cdot \text{magnitude}, \quad (40)$$

$$X_{i2} \leftarrow X_{i2} - 0.5 \cdot \text{magnitude}, \quad (41)$$

$$D_i \leftarrow D_i + 0.4 \cdot \text{magnitude} \cdot \tanh(X_{i3}), \quad (42)$$

$$Y_i \leftarrow Y_i + \text{magnitude} (0.4 \sin(X_{i1}) - 0.3X_{i2}X_{i3} + 0.2|X_{i4}|) + 0.35 \cdot \text{magnitude} \cdot Z_i. \quad (43)$$

In plain language, this means that part of the sample is pushed into a different local pattern. The shift changes some covariates directly, changes how treatment is generated, and changes how the outcome responds as well. So these observations are not just unusual in the way an outlier is unusual. They are generated by a somewhat different process.

That is what makes domain shift different from the earlier contamination mechanisms. Vertical outliers, leverage points, and bad records all create observations that are strange relative to the rest of the sample. Domain shift instead creates a subgroup whose joint distribution looks different. The issue is not simply that a few points are extreme. It is that one part of the data begins to follow a different relationship between the covariates, treatment, and outcome.

This case matters because many applied datasets contain exactly this kind of problem. Data may pool together regions, time periods, institutions, or subpopulations that do not

operate in quite the same way. A model fit to the full sample can then look unstable even when no single observation appears obviously corrupted. Including domain shift in the benchmark therefore makes it possible to study whether robust second-stage procedures still help when the challenge comes from a broader mismatch in the data rather than from isolated contaminated points.

4.3 First-Stage Learners

Because the first-stage learners were introduced earlier, the main task here is not to re-explain them in detail, but to show how they are used in the benchmark. Table 2 summarizes the matched learner design and the main tuning choices used in the simulations.

Table 2: First-stage learners and main hyperparameters used in the simulations.

| Learner | Main hyperparameters |
|-------------------------------|--|
| LassoCV | 3-fold internal CV <code>n_alphas = 25</code> <code>max_iter = 20000</code> |
| ElasticNetCV | 3-fold internal CV <code>n_alphas = 25</code> <code>l1_ratio in {0.1, 0.5, 0.9, 0.95, 1.0}</code> <code>max_iter = 20000</code> |
| RandomForestRegressor | <code>n_estimators = 200</code> <code>min_samples_leaf = 5</code> <code>n_jobs = 1</code> |
| HistGradientBoostingRegressor | <code>max_depth = 6</code> <code>learning_rate = 0.05</code> <code>max_iter = 300</code> <code>min_samples_leaf = 20</code> |

These learners are used in matched fashion rather than averaged over indiscriminately. The sparse-linear design is paired with lasso and elastic net because the true relationships are sparse and linear. The nonlinear and regime-shift designs are paired with random forests and histogram gradient boosting because those learners are better suited to nonlinear patterns in treatment assignment and the outcome. This matters for interpretation. The simulations are meant to compare second-stage estimators, not to manufacture avoidable first-stage failure and then read that as evidence about robustness in the final regression.

4.4 Second-Stage Estimators

The second-stage comparison set is summarized in Table 3. Since these estimators were described conceptually earlier in the paper, the main purpose here is to show how each one is implemented in the benchmark and what type of protection it is meant to provide.

Table 3: Second-stage estimators and main hyperparameters used in the simulations.

| Estimator | Main hyperparameters |
|-----------------|---|
| OLS | |
| Huber | MAD scale estimate Huber weight function $c = 1.345$ |
| MM | Robust pilot Tukey biweight IRLS refinement pilot retains the central 75% of observations by residual size $c = 4.685$ |
| Winsorized | Trimming level 0.90 clips both \tilde{Y} and \tilde{D} at the implied lower and upper quantiles |
| Clipped-score | MAD-based clipping thresholds applied to numerator and denominator score contributions |
| RANSACRegressor | OLS base estimator minimum sample size $\max(10, 0.5n)$ |

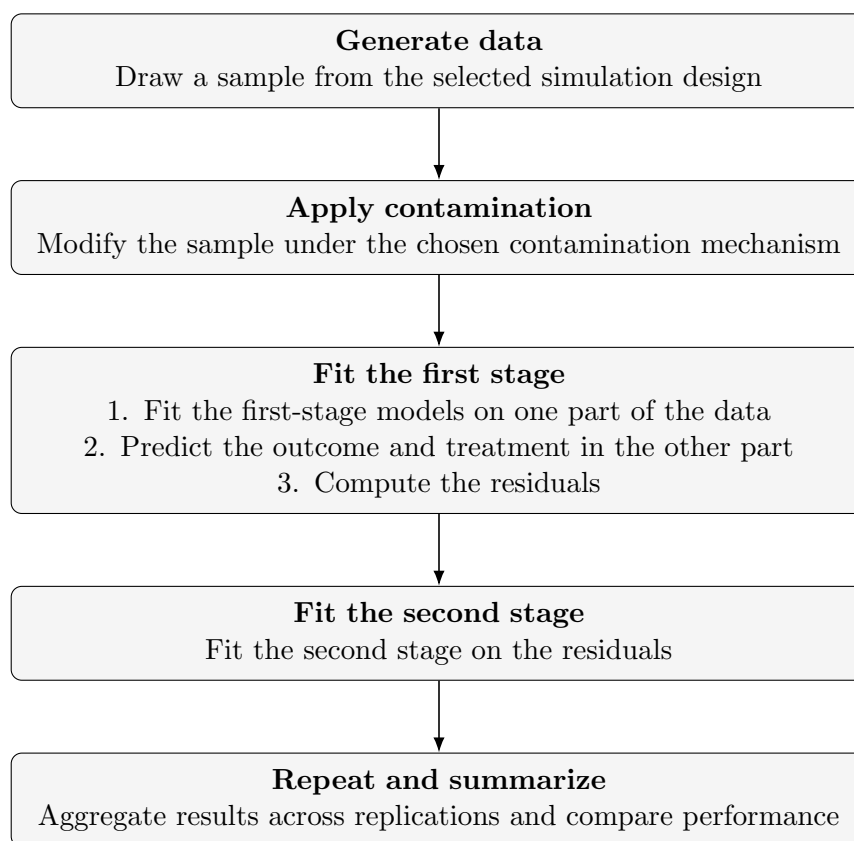
The workflow below already shows how these estimators enter the benchmark procedurally: once the nuisance stage has been cross-fitted, the same residualized sample is passed to each second-stage method. The table here serves a different purpose. It gives the reader a compact reference for what each estimator is, how it is implemented, and why it is included in the comparison.

4.5 Simulation Workflow

Figure 1 summarizes the simulation workflow from data generation through evaluation. The key idea is that each simulation cell fixes a design, generates clean data, applies a chosen contamination mechanism, cross-fits the nuisance stage, and then passes the same residualized sample to each second-stage estimator. This keeps the comparison centered on the final stage rather than on incidental differences in nuisance fitting.

The contamination settings are chosen for the same reason. The simulations consider contamination fractions of 0, 0.01, 0.05, 0.10, and 0.25. The lower end of the grid reflects

Figure 1: Simulation workflow. In each simulation cell, the same first-stage residuals are passed to every second-stage estimator.



a more ordinary reality in social science research: data are rarely pristine. Even when observations are not obviously wrong, survey responses can vary across moments, records can be miscoded or mismatched, and measured quantities can contain small but consequential inconsistencies. The upper end is more severe, but it is included to show how methods begin to separate once those forms of contamination become substantial.

The simulations include seven contamination mechanisms: none, vertical outliers, leverage points, treatment contamination, cellwise contamination, casewise contamination, and domain shift. These mechanisms matter because they do different kinds of damage. Some disturb the residualized outcome, some distort the geometry of the design, some directly corrupt the residualized treatment, and some create a broader mismatch between parts of the sample. A method that handles one of these problems well need not handle the others equally well, so the comparison needs more than a single contamination story.

The second-stage procedures are likewise chosen to reflect methods a researcher might plausibly use. OLS remains the baseline. Huber is the canonical bounded-score alternative. MM is the main method of interest. Winsorized and clipped-score estimators are included because they offer transparent and practically relevant forms of protection, especially in settings where leverage or treatment corruption is the main concern. RANSAC is included because some researchers prefer an explicit inlier-selection approach rather than a score-based robustification.

5 Results

This section examines the performance of the second-stage estimators. Because each grouped scenario computes the residualized sample once and then passes that same residual pair to every second-stage method, the comparisons isolate the second stage rather than differences in nuisance fits. The estimand remains fixed at $\theta_0 = 1$ throughout, so the question is simply how do competing second-stage estimators behave when given the same cross-fitted residualized sample?

The full simulation slate contains 588 grouped scenarios, 3,528 scenario-method cells, and 705,600 method-level results. That scale does not make the conclusions automatic, but it does mean the main patterns are not being driven by a small handful of favorable cells. At the same time, the design remains narrow enough to interpret. The simulations vary the data-generating process, the matched first-stage learner, sample size, contamination type, contamination fraction, and contamination magnitude, but they do so within a structure that keeps the estimand fixed at $\theta_0 = 1$ and keeps the comparison centered on the same

partially linear target throughout.

Beginning with a high level overview, Table 4 summarizes the overall averages across all scenario-method cells. Pooled mean RMSE captures average error magnitude across the full design. Average RMSE rank instead asks where a method tends to place within each simulation cell relative to its competitors. Wins adds a third perspective by counting how often a method finishes first.

Winsorized has the lowest pooled mean RMSE at 0.245 barely ahead of MM at 0.249. Read on its own, that might suggest winsorized is the best overall method. But the rank and win columns qualify that interpretation. MM has by far the best average RMSE rank at 2.301 and records 212 first-place finishes, compared with 119 for Huber and 96 for winsorized. That means winsorized achieves a slightly better pooled average error partly because there are some settings in which it does especially well, enough to pull down its overall mean. MM, by contrast, is the method that most reliably stays near the front across the full design, even when it is not the single best method in any one setting. That matters for how the results should be interpreted. If the goal were to choose a specialist estimator for a setting where leverage or treatment contamination is expected, winsorized would look very attractive. But if the goal is to choose a single robust default for a wide range of unknown applied settings, the rank and win results make MM more compelling.

Table 4: Overall Monte Carlo performance across simulation design cells.

| Method | RMSE | Rank | Wins | Top-2 | Bias | MC 95% interval |
|---------------|-------|-------|------|-------|--------|-----------------|
| MM | 0.249 | 2.301 | 212 | 354 | -0.155 | [0.012, 1.542] |
| Huber | 0.292 | 3.026 | 119 | 296 | -0.176 | [-0.003, 1.718] |
| Winsorized | 0.245 | 3.155 | 96 | 238 | -0.163 | [0.058, 1.512] |
| RANSAC | 0.282 | 3.779 | 7 | 51 | -0.143 | [0.005, 1.757] |
| OLS | 0.382 | 4.357 | 59 | 97 | -0.169 | [-0.139, 2.677] |
| Clipped-score | 0.282 | 4.383 | 95 | 140 | -0.230 | [0.131, 1.376] |

Notes: Results are aggregated at the design-cell level, treating random seeds as replications. Rank, wins, and top-2 finishes are computed within matched cells using RMSE, where lower values are better. Monte Carlo 95% intervals are empirical 2.5th-97.5th percentile ranges of the point estimates. Because all methods use the same first-stage residuals within each design cell, the table compares the second-stage estimators directly rather than evaluating full DML inference.

The overall averages establish the broad pattern of the results, but the more interesting question is how that pattern is built up across the simulations. Diving into the individual contamination mechanisms makes it possible to see where MM’s advantage is most pronounced, where other methods become more competitive, and where all second-stage estimators begin to struggle. The discussion that follows starts with the clean benchmark,

then turns to the main contamination mechanisms, and finally considers how those same patterns vary with sample size, data-generating process, bias, and interval performance.

5.1 Clean-data performance

Table 5: Mean RMSE by contamination type and second-stage method.

| Contamination | MM | OLS | Huber | Winsorized | RANSAC | Clipped-score |
|-------------------------|-------|-------|-------|------------|--------|---------------|
| None | 0.086 | 0.086 | 0.086 | 0.116 | 0.107 | 0.197 |
| Vertical outliers | 0.115 | 0.144 | 0.117 | 0.153 | 0.140 | 0.207 |
| Leverage points | 0.167 | 0.209 | 0.178 | 0.153 | 0.179 | 0.222 |
| Treatment contamination | 0.621 | 0.679 | 0.642 | 0.507 | 0.604 | 0.496 |
| Cellwise | 0.095 | 0.098 | 0.094 | 0.122 | 0.118 | 0.201 |
| Casewise | 0.301 | 0.646 | 0.462 | 0.343 | 0.377 | 0.365 |
| Domain shift | 0.212 | 0.553 | 0.284 | 0.211 | 0.296 | 0.215 |

The clean benchmark asks the most basic practical question, what do we give up by using a robust second-stage estimator when the data are actually well behaved? Table 5 reports results across all contamination settings, including the no-contamination case, which is the setting where OLS should look strongest. Instead, under no contamination, OLS, Huber, and MM are all tied at a mean RMSE of 0.086. That pattern is consistent with a familiar property of MM estimators. When tuned for high efficiency, MM estimators are designed to retain roughly 95% asymptotic efficiency under Gaussian errors, so one would not expect a large clean-data penalty relative to OLS. In these simulations, that intuition appears to carry over to the DML setting. Even under the conditions most favorable to OLS, MM does not sacrifice meaningful point-estimation accuracy for its added robustness.

In the clean benchmark, the methods that pay a clearer price are RANSAC, winsorized, and especially clipped-score. RANSAC is only modestly worse than the leading group, with a mean RMSE of 0.107 rather than 0.086, so its clean-data penalty is real but fairly small. Winsorized does noticeably worse at 0.116, which suggests that its more aggressive handling of extremes can discard useful information when the sample is already well behaved. Clipped-score performs worst by a wide margin at 0.197, indicating that its protection comes with a substantial clean-data cost. The broader pattern, then, is not that these alternatives fail outright, but that in the absence of contamination they buy robustness in a more targeted and more costly way than MM.

5.1.1 Vertical outliers

The vertical-outlier design is the most familiar robust-regression setting in the experiment because the contamination acts directly on the outcome while leaving the covariate structure largely unchanged. In that setting, MM and Huber are the leading methods, with mean RMSE values of 0.115 and 0.117, compared with 0.144 for OLS. For MM, that is about a 20% reduction in RMSE relative to OLS, which is a substantial gain in a design that closely matches the classical motivation for robust regression. This is also an important result for the broader MM-DML argument. MM does not falter even in the canonical setting where Huber would often be the default robust alternative. Huber remains extremely strong here. RANSAC improves somewhat over OLS at 0.140, but it remains well behind MM and Huber. Winsorized and clipped-score are less attractive here, with mean RMSE values of 0.153 and 0.207 respectively. That pattern is consistent with the structure of the problem. In a pure outcome-outlier setting, there is less need for aggressive clipping or trimming, while estimators built to downweight aberrant residuals are better aligned with the contamination being introduced.

5.1.2 Leverage points

The leverage-point design shifts the problem from outcome contamination alone to the geometry of the residualized regression. Here winsorized is the strongest method, with mean RMSE 0.153, followed by MM at 0.167, then Huber and RANSAC at 0.178 and 0.179, while OLS rises to 0.209. Relative to OLS, winsorized lowers RMSE by roughly 27%, and MM also improves meaningfully over the OLS benchmark. MM finishes second overall in this design, trailing winsorized by only 0.014 RMSE, or about 9% relative to the winning value. This is an important qualification to the broader MM-DML argument. MM remains clearly better than OLS, but leverage is one of the clearest settings in which it is not the best performer. Instead, the advantage goes to a method that more directly limits the influence of extreme values.

That pattern is substantively sensible. In the leverage design, the core problem is not just that some responses are unusual, but that a relatively small number of observations are moved into parts of the design space where they can exert disproportionate influence on the fitted second-stage relationship. In that setting, a more direct clipping strategy can outperform residual-based robustification alone.

5.1.3 Treatment contamination

Treatment contamination is where the limits of second-stage robustness become most visible. In this design the regressor of interest is itself corrupted, so the problem is not just that a few observations are extreme, but that the signal entering the second-stage regression has been directly damaged. A robust second-stage estimator can downweight or clip harmful observations, but it cannot fully recover information that has already been lost or distorted in the treatment residual.

The results line up closely with that logic. All methods deteriorate sharply under treatment contamination. OLS is worst at mean RMSE 0.679, but MM and Huber are not far behind at 0.621 and 0.642. The strongest methods here are clipped-score and winsorized, at 0.496 and 0.507 respectively, with RANSAC in between at 0.604. MM finishes fourth overall in this design, improving on OLS but trailing the winning clipped-score method by 0.125 RMSE, or about 25% relative to the winning value. This is an important boundary case for the broader MM-DML argument. The issue is not that MM suddenly fails while another estimator solves the problem. Rather, treatment contamination is simply much harder for every second-stage method because the contamination hits the regressor being used to identify the effect.

That also helps explain why the best-performing methods in this setting are the more aggressive clipping-based approaches. When the treatment residual itself is corrupted, directly limiting the influence of extreme values can help more than residual-based robustification alone. Even so, the winning RMSE values remain high relative to the rest of the design. So the main conclusion is not that some estimator resolves treatment contamination, but that this is one of the clearest settings in which second-stage robustness by itself runs into a fundamental limit.

This result marks an important boundary for the MM-DML story. MM materially improves performance when the issue is bad observations, bad rows, or unstable subgroups. It does not eliminate the consequences of direct corruption in the treatment variable itself. That is a strength of the experiment rather than a weakness. It shows where second-stage robustness helps and where it does not.

5.1.4 Cellwise contamination

Cellwise contamination is a comparatively mild failure mode in this benchmark, and the second-stage results reflect that. Huber is marginally best at mean RMSE 0.094, MM follows at 0.095, and OLS is just behind them at 0.098. These differences are very small. They

suggest that when the contamination is localized to one covariate entry within a row, much of the damage is either absorbed by the first stage or is simply too mild for the second stage to separate strongly.

That result is useful because it helps keep the argument disciplined. The simulations do not imply that second-stage robustification always buys a large gain. In a setting like cellwise contamination, where the corruption is narrow and the residualized regression is not being hit by a large number of severely damaged observations, the gains are modest. This is also one of the cases where clipped-score looks clearly unattractive, with mean RMSE 0.201 and a much larger negative bias than the leading methods. So again the design distinguishes between methods that are broadly useful and methods that achieve protection through a heavier form of shrinkage that becomes costly when the contamination is not severe enough to justify it.

5.1.5 Casewise contamination

Casewise contamination is one of the central results of the experiment because it is the design that most directly mimics badly corrupted observations at the row level. The entire record is damaged, with shifts in covariates, treatment, and outcome together. In that setting MM is the clear leader. Its mean RMSE is 0.301, compared with 0.343 for winsorized, 0.365 for clipped-score, 0.377 for RANSAC, 0.462 for Huber, and 0.646 for OLS. Relative to OLS, MM cuts RMSE by roughly 53%. Relative to Huber, MM reduces RMSE by about 35%. MM also finishes first by a meaningful margin, beating the second-place winsorized method by 0.042 RMSE, or about 14% relative to the MM value.

This is one of the strongest substantive results for the broader MM-DML argument. Casewise contamination is exactly the kind of setting where robust regression needs to handle multiple forms of distortion at once rather than a single isolated problem. Here MM does not merely improve on OLS. It clearly outperforms every alternative in the comparison, including Huber and the more aggressive clipping-based methods. That pattern suggests that when contamination affects whole observations rather than just one margin of the data, the broader protection offered by MM-style robustification can become especially valuable.

5.1.6 Domain shift

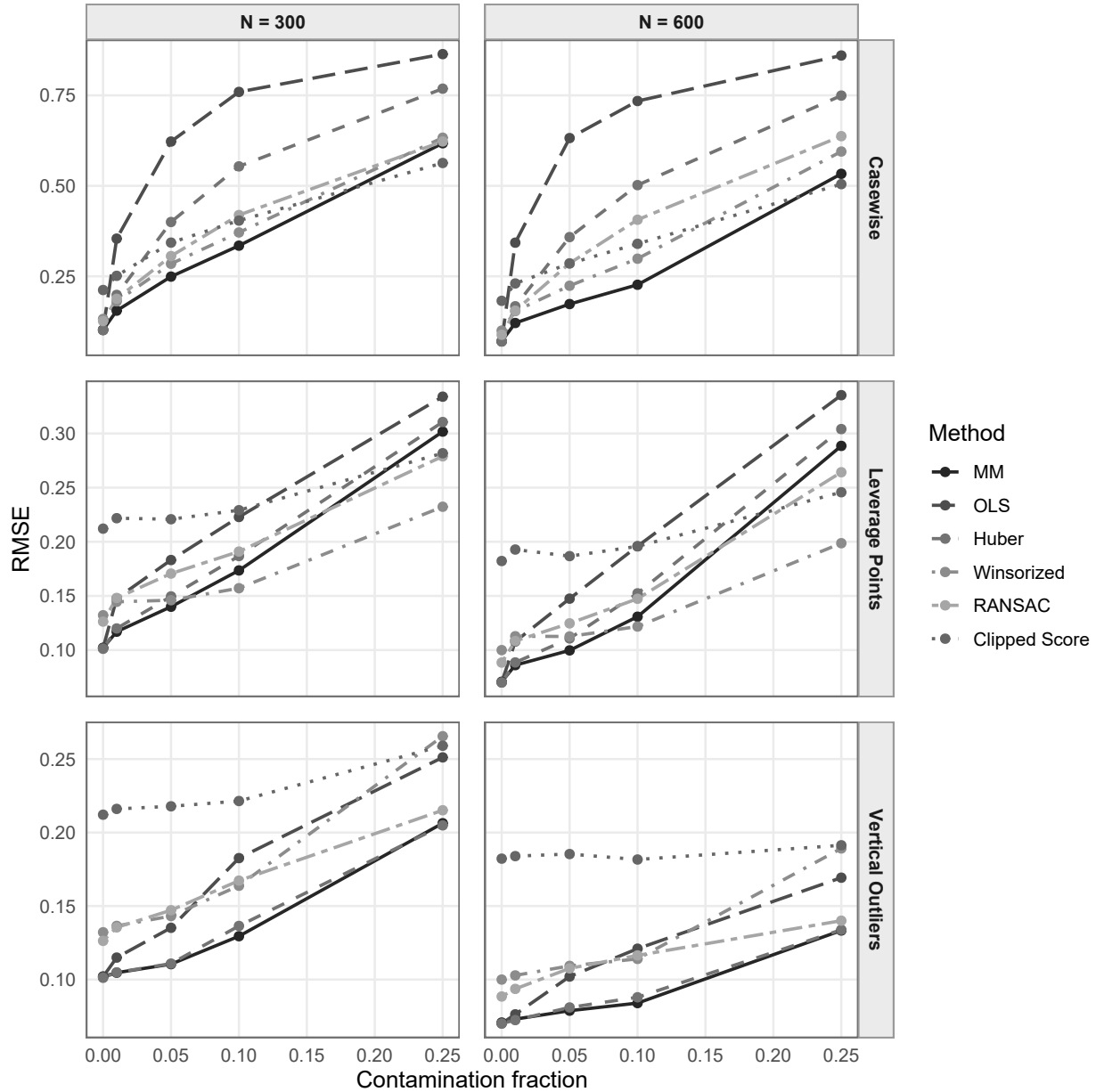
Domain shift is the other central result of the experiment. In this design the problem is not a handful of isolated bad points but a structured subgroup whose relationship between covariates, treatment, and outcome has drifted away from the rest of the sample. That makes it conceptually different from classical outliers, and it turns out to be one of the

hardest settings for OLS.

Under domain shift, OLS has mean RMSE 0.553. MM, winsorized, and clipped-score are all clustered much lower at 0.212, 0.211, and 0.215 respectively. Huber improves on OLS as well, but much less dramatically, at 0.284, while RANSAC reaches 0.296. MM therefore finishes second overall in this design, trailing the winning winsorized method by only 0.001 RMSE, which is negligible in practical terms. Relative to OLS, MM and winsorized each reduce RMSE by about 62%. This is one of the strongest results for the broader MM-DML argument. MM is not uniquely dominant here, but it remains in the very top group in a design where OLS performs extremely poorly.

That pattern is substantively important because it shows that the problem here is not just a few extreme observations that can be downweighted one by one. Instead, part of the sample is behaving as if it comes from a different data-generating process. When OLS tries to fit one squared-error relationship across both the main sample and the shifted subgroup, it can be pulled away from the relationship that describes most of the data, which helps explain why its RMSE becomes so large in this design. By contrast, MM remains essentially tied with the best method, which suggests that its robustness is not limited to classical one-off outliers. It can also help when the data contain a subgroup whose overall pattern does not line up well with the rest of the sample.

Figure 2: RMSE response by sample size



5.2 Sample size and design heterogeneity

The sample-size results are straightforward but more favorable to MM than the pooled averages alone suggest. Moving from $n = 300$ to $n = 600$ reduces the mean RMSE for all methods. MM falls from 0.268 to 0.229, winsorized from 0.265 to 0.226, Huber from 0.309 to 0.274, RANSAC from 0.297 to 0.267, clipped-score from 0.300 to 0.265, and OLS from 0.402 to 0.362. These are meaningful improvements, but they do not materially reorder the

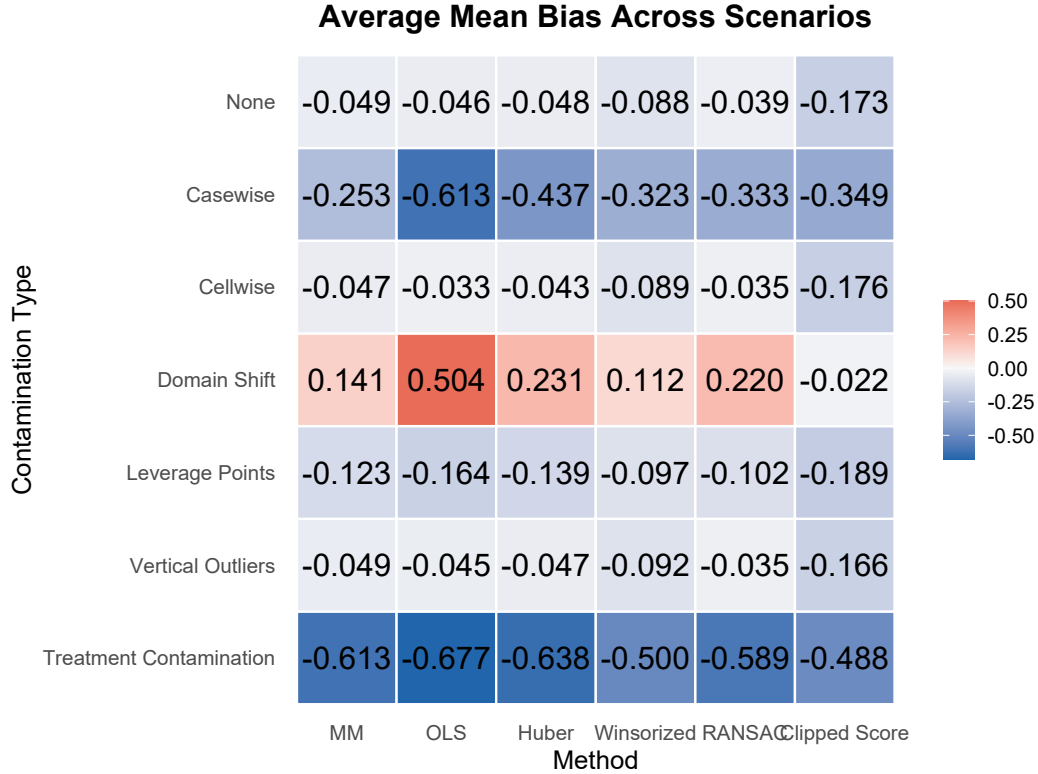
methods. MM has the best average RMSE rank at both sample sizes, with 101 grouped-scenario wins at $n = 300$ and 111 at $n = 600$.

Figure 2 shows why that ranking result is important. MM is not only strong in the cells where it wins outright. Even in the panels where another method has a lower RMSE, MM usually remains close to the frontier until contamination becomes relatively large. This matters because the practical goal of a robust second stage is not necessarily to dominate every specialized competitor in every contamination regime. It is to avoid catastrophic deterioration while preserving good performance across many plausible departures from the clean design. On that criterion, MM looks especially attractive. Even where MM is not the best option, it is still generally competitive.

The improvement from $n = 300$ to $n = 600$ is also substantively meaningful for how MM works. MM estimation begins from a high-breakdown robust fit, then refines that fit with a bounded loss function, here a Tukey-style biweight objective. Observations with small residuals receive approximately ordinary least-squares-like treatment, so the estimator retains high efficiency when the data are clean. Observations with large residuals, however, receive sharply reduced weight, and sufficiently extreme observations can receive essentially no weight in the final estimating equation. In that sense, MM does not simply trim the data mechanically. It uses the fitted residual structure to decide which observations are compatible with the central residualized relationship and which observations should have less influence on the final slope.

With more observations, the residualized relationship is estimated more precisely, so the scale estimate and residual weights become more stable. That helps MM distinguish ordinary noise from observations that are genuinely inconsistent with the main estimating relationship. Put differently, as n grows, the estimator has more clean information with which to identify the dominant signal and can downweight contaminated observations without sacrificing as much precision.

Figure 3: Average Bias



5.3 Bias patterns

Figure 3 is useful because it shows that the methods differ not only in RMSE but also in the direction and structure of their errors. In other words, RMSE tells us how far the estimates are from the truth on average, but the bias results help explain what kind of mistake each method is making. In the simulation designs the true effect is $\theta = 1$, so negative bias means the estimators systematically understate the treatment effect. In practical terms, this pulls estimates toward zero and makes the relationship appear weaker than it actually is. Conversely, when the sign is positive, the models may overstate the relationship between the treatment effect and dependant variable.

First, most methods are already mildly negatively biased even in the clean benchmark. Under no contamination, the average mean bias is about -0.046 for OLS, -0.048 for Huber, and -0.049 for MM. Winsorized is more negative at about -0.088 , and clipped-score is much more negative at about -0.173 . This is important because it shows that robustness is not just a variance story. Some robust procedures also change the center of the estimator in systematic ways. MM is reassuring here because it stays almost exactly in the same range as OLS and Huber in the clean design, while the more aggressively modified estimators show

larger downward bias even before contamination is introduced.

Second, treatment contamination and casewise contamination produce the largest negative biases. Treatment contamination is especially damaging. Every method is pulled well below the true value, with average mean bias ranging from about -0.488 for clipped-score to -0.677 for OLS. MM is also strongly affected at about -0.613 . This is consistent with the broader RMSE results. Once the treatment residual itself is damaged, there is only so much a robust second-stage regression can repair. The estimator is no longer just dealing with unusual outcome values. It is trying to recover a treatment effect from a corrupted treatment signal. Unfortunately, this is the one setting where robust DML methods cannot really save the estimate. When the treatment assignment itself is contaminated, all of the methods break down in similar ways.

Casewise contamination is different, and this is where the bias results are especially helpful for interpreting MM. OLS has a large negative average bias of about -0.613 , while MM is much less biased at about -0.253 . That is not a small difference. In plain terms, MM is doing a much better job recovering the true effect when whole rows of the data are contaminated. Huber also improves over OLS, but remains more biased at about -0.437 . MM is designed to identify a stable central relationship while downweighting observations that are inconsistent with it, and that protection shows up directly in the bias.

Third, domain shift is the one contamination mechanism that flips the sign of the average bias for most methods. OLS is the extreme case, with average mean bias around 0.504 . Huber, MM, RANSAC, and winsorized are also positively biased, though less severely. Clipped-score is the exception, with an average bias near zero at about -0.022 . This sign reversal is substantively important. It suggests that structured subgroup drift behaves differently from ordinary bad-record contamination. The issue is not simply that a few observations are pulling the estimate downward or upward in the usual outlier sense. Rather, part of the sample is following a different local relationship, and OLS responds to that mismatch by shifting too far in the positive direction. MM does not eliminate the problem, but it cuts the bias substantially relative to OLS, from about 0.504 to about 0.141 .

The remaining contamination types are less dramatic but still informative. Under vertical outliers, leverage points, and cellwise contamination, MM stays close to Huber and generally remains less biased than OLS. The differences are not always large, but that is part of the point. MM does not appear to pay a large bias penalty in the easier settings, while it provides much stronger protection in the harder casewise and domain-shift designs.

Together, these bias patterns strengthen the main interpretation of the experiment. The choice of second-stage estimator does not only affect spread. It changes how the esti-

mator responds to different failure modes in the residualized data. OLS can be pushed far from the target when the contamination creates systematic distortion. MM is not magic, especially when the treatment itself is contaminated, but across several of the most important failure modes, it keeps the estimator closer to the true center while performing close to OLS in the clean benchmark.

5.4 What the results imply for MM-DML

Taken together, the results support a fairly clear interpretation of MM-DML. MM is not the best method in every setting. Winsorized is better under leverage and is a strong competitor under treatment contamination. Clipped-score becomes useful in some of the hardest treatment-contamination and domain-shift regimes because its aggressive clipping prevents more dramatic failures. But MM is the method with the most convincing all-around profile.

That profile has three parts. First, MM is essentially costless in the clean benchmark. Second, MM is among the best methods in the classical robust-regression setting of vertical outliers. Third, and most important, MM clearly dominates OLS in the two settings that most directly resemble hard applied-data problems after residualization, namely casewise contamination and domain shift. Those are the settings where a final OLS regression on residuals can become least trustworthy, and those are the settings where MM helps most. This is because once the nuisance stage has been handled through cross-fitting and machine learning adjustments for controls, the residualized regression is still fragile in ways that applied researchers should care about. The results here show that replacing the OLS second stage with MM is a practical and generally effective way to reduce that fragility.

6 Empirical Application

The broader Syria background is well known, and the purpose of this empirical application is not to relitigate that debate. The prewar drought episode is a central part of the case, but the strength of the drought-agriculture-migration-conflict chain remains debated. That makes Syria a useful setting for a methods paper. There is at least a plausible argument that environmental stress could shape local economic and human activity, even if the full drought-migration-conflict story remains contested. Precisely because there are credible arguments on both sides, this is a setting where a more flexible and robust estimator is especially useful. That combination makes this a strong setting for DML with an MM second stage, because the first-stage machine learning can flexibly partial out a rich set of controls while the MM

second stage protects the final residualized regression from the outliers, noisy proxies, and irregular observations that are common in satellite-based measures of environmental and human activity.

6.1 The Drought and Climate Change

While drought is a recurring feature of the region's climate, the 2006–2010 Syria drought is often discussed as unusually severe in both duration and intensity. One influential line of research argues that anthropogenic climate change increased the probability of severe drought in the eastern Mediterranean and helped intensify the Syrian case in particular (Kelley et al., 2015; Hoerling et al., 2012). In this view, the drought is not treated as an isolated weather shock, but as part of a broader regional drying pattern that is becoming more likely under warming conditions.

Drought coincided with longstanding pressures from groundwater depletion, inefficient irrigation, and broader failures of water governance (Gleick, 2014; De Châtel, 2014). That point matters for how the empirical illustration is framed here. The interaction between drought timing and albedo is not meant to isolate a single clean climate mechanism. It is meant to capture whether environmentally stressed places appear to experience a different relationship with local activity during the drought period.

Some accounts push this argument further and suggest that drought, agricultural collapse, migration, and social instability formed part of the background conditions that preceded the Syrian uprising (Kelley et al., 2015; Gleick, 2014). That narrative is important and cannot simply be dismissed, but it is also contested. For a methods paper, that contestation is useful rather than inconvenient. It creates a setting where the substantive story is plausible, the measurement problems are real, and the empirical exercise has to be interpreted with care.

6.2 Quantitative Evidence Refuting the Climate-Migration Narrative

The climate-conflict narrative in Syria has also generated an extensive critical response. A central critique is that the strongest versions of the argument often move too quickly from drought to migration and from migration to civil war. Even if environmental stress harmed rural livelihoods, that does not by itself establish that climate change was a major cause of the conflict.

Selby et al. (2017) argue that the evidence linking anthropogenic climate change to the Syrian drought, large-scale migration, and eventual civil war has often been overstated. Their critique is not that drought was irrelevant, but that the empirical basis for the stronger causal chain remains weaker than many public accounts suggest. De Châtel (2014) similarly argues that focusing too heavily on drought risks obscuring more immediate drivers such as economic liberalization, rural inequality, and policy failures in land and water management. In that reading, climate stress may have mattered, but it mattered through a broader political economy of vulnerability rather than as a stand-alone trigger.

That broader debate shapes how the empirical application is used in this paper. The goal is not to prove that drought caused migration, or that climate change caused the Syrian civil war. The goal is much narrower. The Syria case provides a realistic setting in which environmental measurement is difficult, theory is contested, and applied researchers must still choose variables, justify proxies, and interpret coefficients carefully. That makes it a useful illustration of what MM-DML looks like in practice.

6.3 Measures of Syria’s Changing Environment

To test whether environmentally stressed rural areas experienced a different relationship with local activity during the drought period, the specification uses a small set of core measures capturing local activity, environmental conditions, and drought timing.

Nighttime light intensity serves as the outcome and is used as a proxy for local economic and human activity. These anthropogenic lights reflect population distribution, economic activity, and urbanization patterns. The DMSP-OLS composites provide monthly data from January 2005 to December 2010, allowing the analysis to focus on the prewar period (Elvidge et al., 1997). For spatial consistency, the underlying imagery is binned into $10 \text{ km} \times 10 \text{ km}$ superpixels.

Albedo is the main regressor of interest. It measures the share of incoming light reflected by the surface and is used here as an indicator of terrain and soil conditions. Higher values are associated with more barren or degraded terrain, while lower values are more consistent with vegetated or denser settled areas. The data come from the MODIS MCD43D58 albedo product¹ and provide temporal coverage from January 2005 to December 2010, which makes it possible to track land-surface conditions prior to the onset of civil unrest. In this dataset, albedo ranges from approximately 29.99 to 578.76, with a first quartile of 266.17, a median of 386.41, and a mean of 357.03. Reporting that scale is useful because it clarifies that the estimated coefficients operate over a fairly wide observed range

¹retrieved from <https://search.earthdata.nasa.gov/search>

rather than over a narrowly concentrated variable.

The dichotomous drought variable differentiates the pre-drought and drought phases, with 2006 marking the onset of the national Syria drought period. A key design choice is that drought is defined at the national level rather than through local drought realizations. This is intended to reduce endogeneity arising from local environmental conditions that may jointly influence both treatment assignment and outcomes. The empirical interaction of interest is therefore between albedo and the national drought-period indicator.

The empirical illustration focuses only on rural observations. Using the World Cities Dataset, a 10 km buffer is drawn around each city, and superpixels that intersect that buffer are removed from the sample. This restriction follows the substantive logic of the application. If drought contributed to out-migration, the effect should be most visible in rural areas where agricultural livelihoods were most directly exposed. Centroid coordinates are extracted for each superpixel, not as direct linear controls in the final specification, but to support local spatial modeling and the construction of spatial lags within time slices.

The empirical question is straightforward. In spirit, this interaction design is similar to a continuous-treatment difference-in-differences setup. The drought period acts like a common shock, while albedo captures differences in environmental vulnerability across rural areas. Under a stronger causal reading, the key assumption would be that, absent drought, higher- and lower-albedo rural areas would not have experienced systematically different changes in nighttime lights right when the drought began, and that no other contemporaneous shock differentially affected those areas in the same way. This paper does not claim to fully establish those conditions. Instead, the design is used more modestly to ask whether the relationship between local vulnerability and local activity becomes more negative once the drought period begins. If nighttime lights are interpreted as a proxy for local economic activity or population presence, a negative interaction is consistent with sharper decline in more vulnerable rural areas during drought.

6.4 Design Logic and Estimation Strategy

The Syria dataset is prepared in several steps. A shapefile of Syria is loaded, a 10x10km grid is placed, centroid coordinates are extracted, the drought indicator is converted into a binary drought-period measure, and an interaction term is constructed as $\text{Albedo} \times \text{Drought}$. Month and year are treated as factors, the sample is restricted to rural observations, and the geometry is dropped for estimation. The resulting panel is then augmented with local spatial lags computed separately within each year-month slice using k -nearest-neighbor weights. The local spatial lag of night lights is included as a control. This borrows the core intuition of a

spatial lag model by accounting for local dependence in the outcome, while preserving the flexibility and interpretability of the lasso-based nuisance stage.

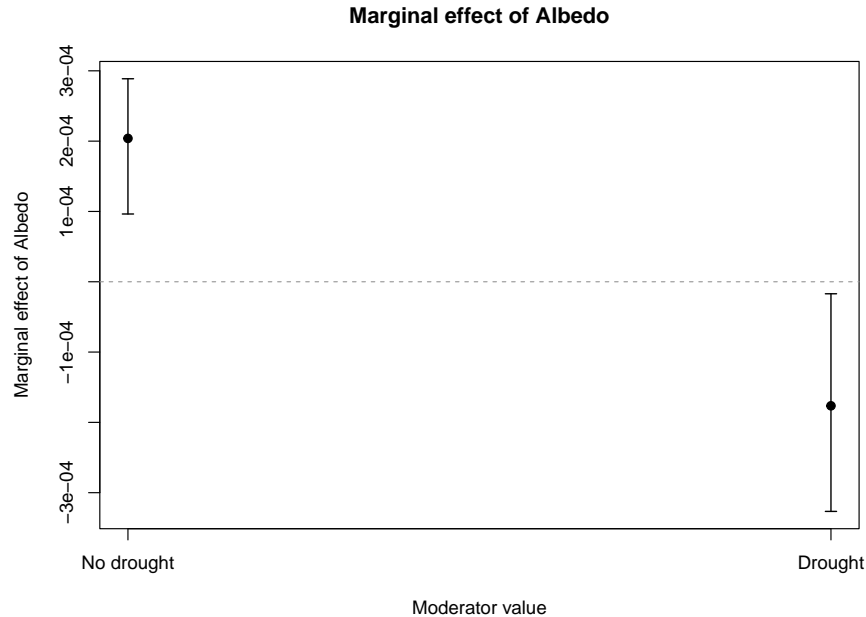
The main empirical MM-DML fit uses lasso in the nuisance stage because it preserves the theory-driven interaction specification and keeps the nuisance stage tied to the target estimand rather than turning the exercise into a pure prediction problem. The fitted specification uses nighttime lights as the outcome, Albedo and Albedo \times Drought as the jointly estimated treatment terms, and Drought, month effects, year effects, and the local spatial lag of night lights as controls. This setup lets the flexible first stage absorb a rich set of adjustments while leaving the MM second stage to estimate the residualized relationship, thereby reducing sensitivity to noisy proxies, outliers, and other irregularities in the data.

6.4.1 Spatial Diagnostics

Because the data are spatial, it is worth checking whether meaningful spatial dependence remains after the nuisance stage. To do that, Moran diagnostics are computed separately within each year-month slice for the orthogonalized outcome residuals, the orthogonalized treatment residuals for Albedo and Albedo \times Drought, and the residuals from the final second-stage MM regression. For the orthogonalized outcome residuals, the median Moran's I across slices is about -0.091 , with the 90th percentile around -0.075 . The residuals from the final second-stage MM regression look nearly identical, with corresponding values of about -0.091 and -0.074 . These diagnostics are not meant to prove that every spatial issue has been solved. They are simply a check on whether obvious positive spatial clustering is still sitting in the variables the estimator ultimately uses. Taken together, these checks suggest that the orthogonalized outcome and treatment variables, along with the residuals from the final second-stage MM fit, are reasonably well-behaved spatially.

6.4.2 Main Results

Figure 4: Marginal Effect of Albedo



The figure plots the marginal albedo effect before drought and during drought, with uncertainty intervals bootstrapped from the joint covariance matrix for the interaction specification.

Figure 4 summarizes the result by plotting the marginal effect of albedo before drought and during drought, with uncertainty intervals bootstrapped from the joint covariance matrix for the interaction specification. The pre-drought effect is given by the coefficient on albedo, while the drought-period effect is the sum of the albedo coefficient and the albedo \times drought interaction. Estimating the two jointly matters because the uncertainty around the drought-period effect depends on both terms together rather than on either coefficient in isolation.

The results show that before the drought, the estimated marginal effect of albedo was positive, but during drought that marginal effect turns negative. In substantive terms, this means that rural areas with surface characteristics associated with greater environmental stress do not appear systematically dimmer before the drought. Once the drought period begins, however, those same areas show a more negative relationship with nighttime lights. In that limited sense, the interaction behaves much like a difference-in-differences style comparison with continuous exposure. The common shock is the drought period, the exposure is environmental vulnerability as proxied by albedo, and the figure suggests that the more exposed rural areas were hit harder once drought conditions set in.

This is not simply a story in which drought lowers activity everywhere in the same way. Rather, the relationship between environmental stress and local activity worsens during drought. That follows the basic logic of the climate-vulnerability argument. The drought did not affect all rural areas equally. Areas with higher albedo, interpreted here as more environmentally stressed or potentially more degraded to begin with, show the stronger negative relationship with nighttime lights once the drought arrives. If nighttime lights are read as a proxy for local economic activity or population presence, one plausible interpretation is that the more vulnerable rural areas experienced a sharper contraction and possibly greater out-migration when the shock finally hit.

The pattern therefore points in the direction of a causal relationship, but establishing that claim is not the purpose of this paper. A stronger causal argument would require at least three additional steps. First, one would need evidence that higher- and lower-albedo rural areas were on comparable pre-drought trends in nighttime lights rather than already moving in different directions. Second, one would need to rule out other shocks arriving at the same time, such as changes in agricultural policy, water access, subsidy regimes, local economic conditions, or early conflict-related disruption, that could have affected higher-albedo areas differently for reasons unrelated to drought itself. Third, one would need stronger confidence that albedo is capturing preexisting environmental vulnerability rather than contemporaneous changes that may themselves reflect migration, land-use change, or economic decline. Those are demanding requirements, and they go beyond the goal of this paper. The point here is narrower. It is to show that MM-DML can recover and present a substantively meaningful interaction effect in a difficult applied setting where the data are noisy, the theory is contested, and the specification is not trivial.

7 Practical Guidance

The benchmark and empirical example point to a fairly practical recommendation for applied work. If the data look clean and there is no strong reason to expect influential residualized observations, OLS-DML remains a reasonable baseline. In the clean benchmark, OLS, Huber, and MM are essentially tied, so there is no reason to say that OLS suddenly becomes a bad estimator in well-behaved settings.

The more important lesson begins once contamination becomes plausible. In most applied social science settings, the analyst does not know in advance whether the main problem is vertical outliers, bad rows, leverage, mild domain shift, or some combination of these. Under that kind of uncertainty, MM-DML is the strongest default second-stage choice

in the benchmark. It stays close to the clean-data frontier, performs very well under vertical outliers and casewise contamination, and remains in the very top group under domain shift. More broadly, MM is near the front in most regimes even when it is not the single best method. That makes it especially attractive in the social sciences, where data are rarely pristine, and the exact failure mode is usually unknown a priori.

That recommendation still needs to be qualified by the likely source of trouble. If leverage-heavy contamination is known to be the main concern, winsorization becomes a serious alternative and can outperform MM. When the treatment residual itself is corrupt, second-stage robustness can only do so much. RANSAC also has a role and may appeal to analysts who care most about conservative interval behavior and are willing to tolerate weaker point estimates and wider intervals in return.

The broader lesson is that orthogonalization and robustness solve different problems, and applied researchers often need both. Orthogonalization reduces sensitivity to nuisance-estimation bias. Robustification reduces sensitivity to influential observations that remain in the residualized regression after the nuisance stage is complete. In messy applied data, those are separate vulnerabilities. Seen that way, MM-DML is not just a small technical modification to standard DML. It is a practical way to harden the final stage of the estimator against the kinds of bad observations that applied researchers routinely encounter but rarely understand perfectly in advance. A simple rule of thumb follows from the benchmark. Keep OLS-DML as a benchmark when the data look ideal, but treat MM-DML as the default once there is any serious reason to think the residualized regression may not be clean.

8 Conclusion

This paper is motivated by a familiar feature of social science data. The substantive relationships researchers care about are often complicated, nonlinear, and highly interactive, which is exactly why double machine learning is so useful. DML makes it possible to flexibly partial out rich nuisance structure without giving up a low-dimensional causal parameter of interest. But the same datasets that make DML attractive are also rarely pristine. They are often noisy, partially subjective, unevenly measured, and vulnerable to influential observations, coding disagreements, subgroup drift, and bad rows. In those settings, the final partially linear regression remains a meaningful point of fragility if it is estimated by ordinary least squares.

There are two related problems. First, even when the original data-generating process is clean, the second stage of practical DML is not estimated on oracle Gaussian errors. It

is estimated on cross-fitted residuals that combine the structural disturbance with first-stage approximation error from both nuisance fits. Because those errors can vary across observations and folds, the residualized regression can contain heteroskedasticity, heavy tails, and influential residualized observations even before any explicit contamination is introduced. Second, the data used in applied social science are rarely clean in the first place. Coding uncertainty, measurement error, outliers, leverage points, bad merges, and subgroup drift can all enter the residualized sample and distort the final OLS regression. Orthogonalization helps protect the target from first-stage regularization bias, but it does not by itself make the final least-squares step robust to these finite-sample and contamination problems.

This paper attempts to fix that weakness of DML. It leaves the standard partially linear DML framework in place and asks a simpler question. What happens if the final OLS regression is replaced with a robust second-stage estimator while everything else is kept as close as possible to the usual design? Framed that way, the contribution is intentionally focused. The goal is not to reinvent DML. It is to make one of its most familiar implementations better suited to the kinds of messy data environments in which social scientists actually work.

The main theoretical result is that under the same orthogonal partially linear structure used by standard DML, MM-DML continues to target the same causal effect and follows the same basic asymptotic logic. That matters because the robustification does not come from changing the estimand. It comes from changing how the final residualized relationship is estimated once the nuisance stage has done its work. In that sense, orthogonalization and robustness solve different problems and should be seen as complementary rather than competing ideas. DML helps when the adjustment problem is complex. MM helps when the residualized data are messy.

The simulation results show why that complementarity matters in practice. In clean data, MM performs essentially as well as OLS, which means robustness does not come with a meaningful clean-data penalty in the main benchmark. Once contamination is introduced, however, the second-stage choice becomes much more important. MM performs especially well under vertical outliers and casewise contamination, remains in the top group under domain shift, and stays broadly competitive across a wide range of bad-data regimes. That pattern is exactly what makes it attractive as a general-purpose default. The benchmark does not show that MM is best everywhere. It shows something more useful for practice. MM is the method that most consistently stays near the front when the analyst does not know in advance what kind of contamination may be present.

That practical point is especially relevant for social science research derived from

difficult observational data. Many widely used datasets include judgment calls, coding uncertainty, partial mismeasurement, or rough category boundaries rather than clean physical measurements. In settings like that, the problem is not always a single dramatic outlier. Sometimes it is a collection of observations that are just noisy enough, inconsistent enough, or influential enough to distort the final residualized regression more than researchers realize. MM-DML is useful precisely because it provides protection against that kind of messiness without requiring the analyst to know in advance exactly where it enters.

That conclusion should still be stated with appropriate limits. Winsorized and clipped-score procedures can outperform MM when leverage or direct treatment corruption is the dominant problem. Treatment contamination in particular marks a real boundary for second-stage robustness. When the treatment residual itself is damaged, no second-stage estimator fully rescues the design. Those are important qualifications, not inconveniences. They help keep the practical recommendation aligned with what the evidence actually supports.

Taken together, the paper's conclusions are practical. Researchers do not need to abandon OLS-DML when the data are clean. But they also should not treat OLS as the automatic default once contamination becomes plausible in the residualized regression. In the kinds of messy but substantively rich datasets that are common in the social sciences, MM-DML offers a strong and broadly reliable alternative. It preserves the familiar partially linear DML structure, takes advantage of DML's flexibility for complex nuisance relationships, and hardens the final stage against the kinds of irregular observations that are common in real data.

Standard DML is valuable because social science relationships are often complex. MM-DML is valuable because social science data are often messy. Many of our applications are both complex and messy, and in such settings, orthogonalization alone solves only half the problem. If the data are complex, noisy, inconsistently coded, or vulnerable to influential observations, researchers should not stop at traditional DML. They should use MM-DML instead. It preserves DML's flexibility for nonlinear and interactive adjustment while making the final stage less fragile in exactly the kinds of data environments that are common in applied social science.

References

- Andersen, R. (2008). *Modern Methods for Robust Regression*. Sage, Thousand Oaks, CA.
- Baissa, D. K. and Rainey, C. (2020). When blue is not best: non-normal errors and the linear model. *Political Science Research and Methods*, 8(1):136–148.
- Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.
- De Châtel, F. (2014). The role of drought and climate change in the syrian uprising: Untangling the triggers of the revolution. *Middle Eastern Studies*, 50(4):521–535.
- Dixon, W. J. and Yuen, K. K. (1974). Trimming and winsorization: A review. *Statistische Hefte*, 15(2):157–170.
- Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., Davis, E. R., and Davis, C. W. (1997). Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. *International Journal of Remote Sensing*, 18(6):1373–1379.
- Freedman, D. A. (2006). On the so-called “huber sandwich estimator” and “robust standard errors”. *The American Statistician*, 60(4):299–302.
- Gleick, P. H. (2014). Water, drought, climate change, and conflict in syria. *Weather, climate, and society*, 6(3):331–340.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.
- Harada, K. and Fujisawa, H. (2024). Outlier-resistant estimators for average treatment effect in causal inference. *Statistica Sinica*, 34:133–155.
- Hoerling, M., Eischeid, J., Perlwitz, J., Quan, X., Zhang, T., and Pegion, P. (2012). On the increased frequency of mediterranean drought. *Journal of climate*, 25(6):2146–2161.
- Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

- Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821.
- Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley, Hoboken, NJ, 2 edition.
- Kelley, C. P., Mohtadi, S., Cane, M. A., Seager, R., and Kushnir, Y. (2015). Climate change in the fertile crescent and implications of the recent syrian drought. *Proceedings of the national Academy of Sciences*, 112(11):3241–3246.
- King, G. and Roberts, M. E. (2015). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, 23(2):159–179.
- Krueger, J. S. and Lewis-Beck, M. S. (2008). Is ols dead? *The Political Methodologist*, 15(2):2–4.
- Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics, Volume 4*, pages 2111–2245. Elsevier, Amsterdam.
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880.
- Rousseeuw, P. J. and Yohai, V. (1984). Robust regression by means of s-estimators. In Franke, J., Härdle, W., and Martin, D., editors, *Robust and Nonlinear Time Series Analysis*, volume 26 of *Lecture Notes in Statistics*, pages 256–272. Springer, New York.
- Selby, J., Dahi, O. S., Fröhlich, C., and Hulme, M. (2017). Climate change and the syrian civil war revisited. *Political Geography*, 60:232–244.
- Tukey, J. W. and McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/winsorization 1. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 331–352.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.
- Wang, X., Liu, Y., Qin, G., and Yu, Y. (2024). Robust double machine learning model with application to omics data. *BMC Bioinformatics*, 25(1):355.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.
- Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2):642–656.

A Causal Interpretation and Formal Theorem

This appendix explains why MM-DML can still be interpreted as a causal estimator in the partially linear model and then records the final theorem stack in paper notation. The contribution is not inventing orthogonality or MM estimation, but showing that a robust MM-style second-stage score can be embedded into the partially linear DML framework while preserving the same target parameter and a corresponding formal large-sample theory under explicit conditions. The central point is narrow but important. Robustification changes the second-stage score, but it does not change the target parameter. Under the same partially linear identifying restrictions used by standard DML, the robust score remains centered at the same causal effect, remains locally identifying in the target direction, and remains orthogonal to first-order nuisance perturbations. Cross-fitting then does the same work it does in standard DML: it prevents first-order nuisance estimation error from dominating the low-dimensional target equation. What the robust score adds is resistance to contaminated residual contributions in that final equation.

Two scope notes matter. First, the formal development is strongest on score construction, identification, orthogonality, estimator packaging, consistency, and asymptotic linearity. Second, the asymptotic-normality and inference layers are formalized through proxy and event-lifting arguments rather than through a fully internalized central limit theorem. That boundary should remain explicit because it is the accurate description of what is machine checked.

A.1 Mathematical Roadmap

Write the partially linear model as

$$Y = \theta_0 D + g_0(X) + U, \quad D = m_0(X) + V,$$

with

$$E[U | X, D] = 0, \quad E[V | X] = 0.$$

Define the residualized treatment and structural residual as

$$V = D - m_0(X), \quad U = Y - \theta_0 D - g_0(X).$$

The population robust score is

$$\Psi(\theta, g, m, s) = E \left[(D - m(X)) \psi \left(\frac{Y - g(X) - \theta(D - m(X))}{s} \right) \right].$$

At the truth,

$$\Psi(\theta_0, g_0, m_0, s_0) = E \left[V \psi \left(\frac{U}{s_0} \right) \right].$$

Under the centering condition used in the main text, this expectation is zero. That is the first step in the causal argument: the robust score is centered at the same structural parameter that appears in the partially linear model, so the estimator is not pursuing a different estimand.

The second step is local identification. Differentiating the population score with respect to θ at the truth yields the Jacobian

$$J := \partial_\theta \Psi(\theta, g_0, m_0, s_0) \Big|_{\theta=\theta_0},$$

which is nonzero under the paper's nondegeneracy condition. Intuitively, once the residualized treatment still contains variation after conditioning on X , movement in θ changes the population score locally. That makes θ_0 an isolated solution to the estimating equation rather than merely one value among many.

The third step is Neyman orthogonality. The score is constructed so that first-order perturbations in the nuisance functions g and m vanish at the truth. Econometrically, the implication is the familiar DML one: small nuisance-estimation errors do not enter the target equation at first order. That is what makes it possible to use flexible first-stage learners without immediately giving up the low-dimensional treatment effect.

The sample estimator is defined from the cross-fitted residuals

$$\tilde{Y}_i = Y_i - \hat{g}^{(-k(i))}(X_i), \quad \tilde{D}_i = D_i - \hat{m}^{(-k(i))}(X_i),$$

through the empirical score

$$\hat{\Psi}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \tilde{D}_i \psi \left(\frac{\tilde{Y}_i - \theta \tilde{D}_i}{\hat{s}} \right).$$

The MM-DML estimator is any local approximate root satisfying

$$|\hat{\Psi}_n(\hat{\theta}_n)| \leq \text{tol}_n.$$

The theorem stack then follows the standard DML route plus robust-score regularity. Score centering and nondegeneracy identify θ_0 . Orthogonality suppresses first-order nuisance error. Cross-fitting converts empirical score fluctuations into a leading term plus a higher-order remainder. Asymptotic linearity then delivers the estimator’s large-sample behavior, and proxy-based Gaussian approximation yields the corresponding inference layer.

The substantive takeaway is straightforward. MM-DML is not merely a robust regression fit applied after residualization. Under the partially linear causal model and the stated regularity conditions, it targets the same causal parameter as standard DML while using a second-stage score that is less sensitive to contaminated residuals.

A.2 Paper-Facing Statements Corresponding to the Formal Stack

The final theorem stack establishes five core claims in sequence:

- (i) $\Psi(\theta_0, g_0, m_0, s_0) = 0$,
- (ii) $J := \partial_\theta \Psi(\theta, g_0, m_0, s_0)|_{\theta=\theta_0} \neq 0$,
- (iii) $\partial_g \Psi(\theta_0, g_0, m_0, s_0)[h] = 0$ and $\partial_m \Psi(\theta_0, g_0, m_0, s_0)[q] = 0$,
- (iv) $\hat{\theta}_n \xrightarrow{p} \theta_0$,
- (v) $\sqrt{n}(\hat{\theta}_n - \theta_0) = J^{-1}L_n + o_p(1)$,

followed by proxy-based Gaussian approximation and proxy-valid Wald inference for the studentized estimator.

Exact packaged population object

The Lean presentation layer packages the population result as

$$\text{PopulationCoreResult}(\mu, M),$$

with five fields.

`score_zero`

$$\text{popScore}(\mu, M, M.\theta_0, M.l_0, M.m_0, M.s_0) = 0.$$

theta_hasDerivAt

$$\text{HasDerivAt} \left(\begin{array}{l} \theta \mapsto \text{popScore}(\mu, M, \theta, M.\ell_0, M.m_0, M.s_0), \\ \int \text{thetaScoreDerivPoint}(M, \omega) d\mu, \\ M.\theta_0 \end{array} \right).$$

theta_deriv_nonzero

$$\int \text{thetaScoreDerivPoint}(M, \omega) d\mu \neq 0.$$

ell_orthogonal

$$\forall h, \text{HasDerivAt} \left(\begin{array}{l} t \mapsto \text{popScore}(\mu, M, M.\theta_0, \text{ellPerturb}(M, h, t), M.m_0, M.s_0), \\ 0, \\ 0 \end{array} \right).$$

m_orthogonal

$$\forall q, \text{HasDerivAt} \left(\begin{array}{l} t \mapsto \text{popScore}(\mu, M, M.\theta_0, M.\ell_0, \text{mPerturb}(M, q, t), M.s_0), \\ 0, \\ 0 \end{array} \right).$$

Named assumption bundles

The Lean files also make the assumption architecture explicit. The population theorem is stated through $\text{PopulationAssumptions}(\mu, M)$, the primitive bridge through $\text{PrimitiveCausalAssumptions}(\mu, E)$ and $\text{AnalyticAssumptions}(\mu, M)$, the consistency theorem through $\text{PrimitiveScoreApproxAssumptions}(\mu, E)$, the linearization theorem through $\text{PrimitiveLinearizationAssumptions}(\mu, E, \theta_0, J)$, the leading-term Gaussian bridge through $\text{PrimitiveLeadingProxyAssumptions}(\mu, E, \theta_0, G)$, and the studentization bridge through $\text{PrimitiveStudentizationAssumptions}(\mu, E, \theta_0, \hat{\tau}, \tau_0, G)$.

Theorem A.1 (Population identification and orthogonality). *Let μ be a measure and let M be a formal probability model.*

Assume

$$\begin{array}{l} \text{PopulationAssumptions}(\mu, M), \\ \int \text{thetaScoreDerivPoint}(M, \omega) d\mu \neq 0. \end{array}$$

Then

$$\text{PopulationCoreResult}(\mu, M).$$

Equivalently, the population score is zero at the true parameter, differentiable in the target direction at $M.\theta_0$ with derivative

$$\int \text{thetaScoreDerivPoint}(M, \omega) d\mu,$$

that derivative is nonzero, and the first-order derivatives in the ℓ and m nuisance perturbation directions vanish at the population level.

Theorem A.2 (Primitive assumptions imply the population score conditions). *Let μ be a measure and let M be a formal probability model.*

Assume

$$\begin{aligned} & \text{PrimitiveCausalAssumptions}(\mu, M), \\ & \text{AnalyticAssumptions}(\mu, M), \\ & \int \text{thetaScoreDerivPoint}(M, \omega) d\mu \neq 0. \end{aligned}$$

Then

$$\text{PopulationCoreResult}(\mu, M).$$

In the formal development, this theorem is proved by first constructing

$$\text{PopulationAssumptions}(\mu, M)$$

from the primitive and analytic bundles and then invoking Theorem A.1.

Definition A.3 (Cross-fitted MM-DML estimator). The paper-facing Lean development uses the bundled object

$$\text{CrossFitMMEstimator}(n, K, \text{Cov}),$$

re-exported as

$$\text{CrossFittedMMDMLEstimator}(n, K, \text{Cov}).$$

A bundled cross-fitted MM-DML estimator consists of fields

$$\text{inputs}, \quad \Theta, \quad \text{tol}, \quad \hat{\theta},$$

together with the properties

$$\hat{\theta} \in \Theta, \quad \left| \text{empiricalScore}(\text{inputs}, \hat{\theta}) \right| \leq \text{tol}.$$

Setting $\text{tol} = 0$ recovers the exact-root case.

Proposition A.4 (Bundled estimator implies approximate-root property). *Let E be a bundled cross-fitted MM-DML estimator. Then*

$$\text{IsApproxRootOn}(E.\text{inputs}, E.\Theta, E.\text{tol}, E.\widehat{\theta}).$$

That is, every bundled estimator is an approximate root on its declared parameter set by construction.

Theorem A.5 (Consistency from primitive rate conditions). *Let μ be a sequence of measures, let E_n be a sequence of bundled cross-fitted MM-DML estimators, let $\text{popScore} : \mathbb{R} \rightarrow \mathbb{R}$, and let $\theta_0 \in \mathbb{R}$. If*

$$\text{PrimitiveScoreApproxAssumptions}(\mu, E, \text{popScore}, \theta_0)$$

holds, then

$$\text{ConsistentInProbability}(\mu, (n, \omega) \mapsto (E_n(\omega)).\widehat{\theta}, \theta_0).$$

Theorem A.6 (Asymptotic linearity from primitive linearization rates). *Let μ be a sequence of measures, let E_n be a sequence of bundled cross-fitted MM-DML estimators, let $\theta_0 \in \mathbb{R}$, and let $J \in \mathbb{R}$. If*

$$\text{PrimitiveLinearizationAssumptions}(\mu, E, \theta_0, J)$$

holds, then

$$\text{MMDMLAsymptoticallyLinearInProbability}(\mu, n\text{Size}, k\text{Size}, E, \theta_0, J).$$

Theorem A.7 (Supporting bridge: Leading-term Gaussianity via primitive proxy assumptions). *If*

$$\text{PrimitiveLeadingProxyAssumptions}(\mu, E, \theta_0, G)$$

holds, then

$$\text{MMDMLLeadingAsymptoticallyGaussianViaProxy}(\mu, n\text{Size}, k\text{Size}, E, \theta_0, G).$$

This is a proxy-based wrapper, not a fully internal central limit theorem.

Theorem A.8 (Estimator-level asymptotic normality via a scaled proxy). *The fifth Lean theorem is stated through an explicit high-probability event-lifting hypothesis. For every $\varepsilon, \eta > 0$, there exists N such that for all $n \geq N$, there exists a measurable event A with probability*

at least $1 - \eta$ and constants $\text{linTube}, \text{leadTube} \in \mathbb{R}$ such that for every $\omega \in A$,

$$\begin{aligned} & \left| \text{scaledError}(n\text{Size } n, (E_n(\omega)).\widehat{\theta}, \theta_0) - J^{-1} \text{mmDMLLeadingTerm}((E_n(\omega)).\mathbf{inputs}, \theta_0) \right| < \text{linTube}, \\ & |\text{mmDMLLeadingTerm}((E_n(\omega)).\mathbf{inputs}, \theta_0) - G_n(\omega)| < \text{leadTube}, \\ & \text{linTube} + |J^{-1}| \text{leadTube} < \varepsilon. \end{aligned}$$

Under that hypothesis,

$$\text{MMDMLEstimatorAsymptoticallyGaussianViaProxy}(\mu, n\text{Size}, k\text{Size}, E, \theta_0, J, G).$$

Equivalently, the scaled MM-DML estimator is asymptotically Gaussian via the scaled proxy $J^{-1}G$. This remains a proxy/event-lifting theorem rather than a fully internalized central limit theorem.

Theorem A.9 (Supporting bridge: Studentized Gaussianity from primitive joint assumptions). *If*

$$\text{PrimitiveStudentizationAssumptions}(\mu, E, \theta_0, \widehat{\tau}, \tau_0, G)$$

holds, then

$$\text{StudentizedAsymptoticallyGaussianViaProxy}(\mu, n\text{Size}, k\text{Size}, E, \theta_0, \widehat{\tau}, \tau_0, G).$$

Equivalently, the studentized MM-DML statistic is asymptotically Gaussian via the proxy $\tau_0 G$.

Corollary A.10 (Valid inference via proxy). *The final Lean corollary is stated through an event-lifting proxy-control hypothesis. For every $\eta > 0$, there exists N such that for all $n \geq N$, there exists an event A with probability at least $1 - \eta$ such that for every $\omega \in A$,*

$$\begin{aligned} & \left| \widehat{\tau}_n(\omega) \text{scaledError}(n\text{Size } n, (E_n(\omega)).\widehat{\theta}, \theta_0) - \text{proxy}_n(\omega) \right| < \varepsilon, \\ & |\text{proxy}_n(\omega)| \leq c - \varepsilon. \end{aligned}$$

Under that hypothesis,

$$\text{AsymptoticallyValidWaldViaProxy}(\mu, n\text{Size}, k\text{Size}, E, \theta_0, \widehat{\tau}, \text{proxy}, c).$$

That is, the Wald event is asymptotically valid via proxy control.

Bottom line. Stated literally, the checked development proves the following stack in paper order: a packaged population-core result with exact fields for score-zero, target differentiability, nonzero Jacobian, and nuisance orthogonality; a primitive-to-population bridge from `PrimitiveCausalAssumptions` and `AnalyticAssumptions`; a bundled cross-fitted estimator object together with its built-in approximate-root property; consistency in probability from `PrimitiveScoreApproxAssumptions`; asymptotic linearity in probability from `PrimitiveLinearizationAssumptions`; leading-term Gaussianity via proxy from `PrimitiveLeadingProxy`; estimator-level Gaussianity via a scaled proxy under an explicit good-event lifting hypothesis; studentized Gaussianity via proxy from `PrimitiveStudentizationAssumptions`; and asymptotically valid Wald inference via an explicit proxy-control good-event hypothesis.

What the formal development does *not* claim to do is prove a fully internal central limit theorem from primitive empirical-process assumptions, or derive the analytic bridge from a full internal theory of differentiation under the integral sign. Those boundaries are part of the formal theorem design itself.