OPERATOR-ROBUST INSTRUMENTAL VARIABLE REGRESSION UNDER JOINT DISTRIBUTION SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Instrumental variable regression identifies causal effects under unobserved confounding, but only when moment conditions hold. However, weak identification increases the sampling error, while invalid instruments and distribution shifts can violate its moment restrictions. Instead of indirectly protecting against perturbations to the data distribution, we directly robustify the identifying moments. By tracing joint distributional perturbations through these moments, we derive a common score that jointly perturbs the treatment operator and outcome moment while preserving their relationship. This pushforward construction creates an operator-robust worst-case criterion and tractable convex average-case surrogates. A special case connects to k-class regularization under whitening and constant leverage. We characterize how score robustness changes identification and when it preserves the causal target, and bound structural error in terms of identification strength, estimation error, and robustness-induced biases. In the studied synthetic and semi-synthetic settings, the criteria improve structural recovery in several weak-instrument and shifted designs. On real datasets, we assess held-out moment fit because structural ground truth is unavailable.

Code— https://anonymous.4open.science/r/628B

1 INTRODUCTION

Learning true causal relationships from observational data is a fundamental challenge in machine learning, particularly because unobserved confounding factors often simultaneously influence both the intervention and the outcome. Instrumental Variable (IV) regression is a principal method to overcome this. By leveraging an instrument variable Z that affects the treatment T but has no direct effect on the outcome Y nor shares unobserved confounders with it, IV methods can isolate the unconfounded causal effect of T on Y. For a treatment T, outcome Y, and instrument Z, we seek a structural function f that satisfies the Conditional Moment Restriction (CMR)

$$\mathbb{E}_{P_0}[Y - f(T) \mid Z] = 0\,,$$

where P_0 denotes the population distribution (Chamberlain, 1987; Newey & Powell, 2003).

The CMR asks the residual to have zero mean at almost every value of Z, which is difficult to check directly. It implies the unconditional moments $\mathbb{E}_{P_0}[(Y - f(T))h(Z)] = 0$ for every integrable test function h. Modern methods test a candidate f with a flexible class $\mathcal{H}$ of such functions. A restricted class detects only the violations visible through its tests. For instance, maximum moment restriction (MMR) selects a reproducing kernel Hilbert space (RKHS) to measure the largest moment violation (Zhang et al., 2023), while minimax estimators learn the detecting function jointly with the structural model (Dikkala et al., 2020).

However, under weak identification, different structural functions can produce nearly indistinguishable moment violations, heavily inflating the sampling error. Furthermore, invalid instruments and distribution shifts can violate the underlying moment restrictions. To address this, recent methods seek causal models by searching for perturbations of the empirical data distribution which yield a valid model. For instance, they use divergence-based reweighting, kernel measures, and transport metrics to perturb distributions in the search of valid models (Kremer et al., 2022; 2023; Kremer & Schölkopf, 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, some perturbations change both sides of the moment equation which identifies the model f, but *not every change has the same causal meaning*. For example, shifting values of Z while keeping $P(T, Y \mid Z)$ intact preserves a valid CMR, whereas shifting on Y changes the conditional relationship and invalidate it. So, which changes should an IV estimator control, and what does that choice cost? In other words, while instrument-only directions preserve the CMR and can be controlled without moving the population target, general joint directions include local violations.

We construct IV estimators that control the response of the identifying moments to joint distributional changes. Tracing these changes through the moments leades to a joint ambiguity set of *treatment operators* and *outcome moments* while preserving the coupling induced by their common distribution. Then, we compare scores that depend only on instruments with scores that describe general joint changes. For worst-case robustness, we maximize over this set, obtaining an operator-robust estimator. For computational efficiency, we then derive convex criteria. In summary, our main contributions are

- We construct IV estimators from joint treatment-operator and outcome-moment responses.
- We envelop the worst-case through convex lower and upper bounds, yielding a family of convex estimators.
- We show that our score directions decompose into those directions that preserve the CMR and instrument violations.
- We characterize how our our estimators either preserve or shift the causal target.
- We evaluate accuracy on synthetic and semi-synthetic designs and real-world real datasets, including with weak-IV, perturbed, and shifted settings.

2 Preliminaries

We build on conditional moment embeddings and maximum moment restriction (Muandet et al., 2020; Zhang et al., 2023). For observations $\mathcal{D} = \{(t_i, y_i, z_i)\}_{i=1}^n$ sampled from P_0, we assume that there is a true function f_0 that satisfies the CMR $\mathbb{E}_{P_0}[Y - f_0(T) \mid Z] = 0$. At the truth, the residuals $\psi(T, Y; f) = Y - f(T)$ vanish, multiplying with functions on Z will not change this. Thus, we can equivalently restrict the product $\mathbb{E}_{P_0}[\psi(T, Y; f_0)h(Z)] = 0$ for a family of 'test' functiosn h.

We choose an instrument RKHS $\mathcal{H}$ for h with kernel k to test these moments. Using its reproducing property, we obtain the *conditional moment embedding* (Muandet et al., 2020)

$$\mu_P(f) = \mathbb{E}_P[\psi(T, Y; f)k(Z, \cdot)] \in \mathcal{H}, \qquad \mathbb{E}_P[\psi(T, Y; f)h(Z)] = \langle h, \mu_P(f)\rangle_{\mathcal{H}} \,. \tag{1}$$

To prevent arbitrary rescaling, we restrict test functions to the unit ball, obtaining the maximum moment restriction (Muandet et al., 2020; Zhang et al., 2023)

$$\mathcal{L}_P(f) := \sup_{\|h\|_{\mathcal{H}} \leq 1} |\mathbb{E}_P[\psi(T, Y; f)h(Z)]| = \|\mu_P(f)\|_{\mathcal{H}}. \tag{2}$$

We want to track how a shift changes both the outcome moments and the treatment terms that must explain them. For our causal model class, we, thus, choose an RKHS $\mathcal{F}$ with kernel ℓ. Substituting its representation $f(T) = \langle f, \ell(T, \cdot)\rangle_{\mathcal{F}}$ into the moment embedding creates a simple affine function

$$\mu_P(f) = b_P - A_P f \tag{3}$$

$$\text{where} \qquad b_P = \mathbb{E}_P[Yk(Z, \cdot)] \in \mathcal{H}, \qquad A_P = \mathbb{E}_P[k(Z, \cdot) \otimes \ell(T, \cdot)] : \mathcal{F} \to \mathcal{H} \,, \tag{4}$$

where $u \otimes v$ maps f to $u\langle v, f\rangle$ which allows us to treat f as a linear variable that we can pull out of the expectation, leading to the operator form. The treatment operator A_P maps a structural function to its instrument moments, while b_P contains the outcome moments it must match.

3 Operator-Robust Moment Estimation

A satisfied CMR implies $b_P = A_P f$ for a valid structural function, provided the moments exist. However, this equation depends on the joint distribution of treatments, outcomes, and instruments. That means, changing the distribution, therefore, may change both sides of the equation, and a function that satisfies the reference moments may violate the shifted moments.

To study how these conencted changes affects estimation, we start by introducing shifts. Using *divergence-based distributional robustness* (Duchi et al., 2021), we consider shifts that change the probability of data relative to a reference distribution P. We describe this relative change using a weight function $a(T, Y, Z)$, which increases (positive) or decreases (negative values) the relative importance for a joint (T, Y, Z) observation. Through a score-based approach (Duchi et al., 2021), the resulting distribution is defined relative to the reference as

$$dQ = (1 + a)\, dP, \qquad \mathbb{E}_P[a] = 0, \qquad 1 + a \geq 0\ . \tag{5}$$

Here, we measure the strengths of the shift by its norm $\|a\|_{L^2(P)}$, whose square equals the Pearson χ^2 divergence $\chi^2(Q\|P)$ Duchi et al. (2021). Next, to see how such a shift affects identification, we write the moment contribution of one observation $V = (T, Y, Z)$ as a function $\xi_f(V) = [Y - f(T)]k(Z, \cdot)$. Since the shifted distribution Q weights each observation by $1 + a$ in equation 5, its moments are weighted P-moments, $\mu_Q(f) = \mathbb{E}_Q[\xi_f] = \mathbb{E}_P[(1 + a)\xi_f]$, centering $\mu_P(f) = \mathbb{E}_P[\xi_f]$ yields to

$$\mu_Q(f) - \mu_P(f) = \mathbb{E}_P[a(V)\xi_f(V)] =: \mathcal{D}_f a\ . \tag{6}$$

Because each reweighting changes both A_P and b_P, we maintain their relationship. Formally, this brings us to the *pushforward* argument. That is, we can interpret Q as a map $Q \mapsto (A_Q, b_Q)$ that turns a distributional ambiguity set $\mathcal{P}_\rho$ to the *joint moment ambiguity set* $\mathcal{U}_\rho = \{(A_Q, b_Q) : Q \in \mathcal{P}_\rho\}$, such that

$$\sup_{Q \in \mathcal{P}_\rho} \|\mu_Q(f)\|_{\mathcal{H}} = \sup_{(A,b) \in \mathcal{U}_\rho} \|b - Af\|_{\mathcal{H}}\ . \tag{7}$$

In words, the same value a shifts $A_Q = A_P + \mathbb{E}_P[a\, k(Z, \cdot) \otimes \ell(T, \cdot)]$ and $b_Q = b_P + \mathbb{E}_P[aY k(Z, \cdot)]$ jointly, thereby creating a 'structured' uncertainty set for robust least squares problems akin to El Ghaoui & Lebret (1997).

While standard methods seek the smallest shifts under which the CMR is satsifed and moments approach zero (Newey & Smith, 2004; Kremer et al., 2022), we, in contrast, minimize against the *largest moment violation* over the the ambiguity set equation 7, where small losses guarantee that the candidate approximately satisfies the identifying equation under every admissible shift. However, this guarantee depends on which distributional changes we ask the estimator to control.

First, we consider changes only in the Z distribution to keep $P(T, Y \mid Z)$ and CMR intact. This motivates the *kernel-sieve model* (KS), whose score space $\mathcal{S}$ is spanned by centered features of Z. Second, we consider changes in the joint distribution of (T, Y, Z) which can also alter the treatment–outcome relationship. For this, we study the full *tangent model* (TAN) allows every admissible score. As the broader 'TAN' includes directions that can break the CMR even at f_0, we may pull the estimator away from the structural target. To interpret a large moment violation, we need to know whether it reflects a poor candidate or a shift that invalidates the CMR even at f_0. For the tangent model we are motived by the conditional-moment efficiency theory (Chamberlain, 1987; Newey, 1993).

Proposition 1 (Preserving and violating directions) *Suppose $\varepsilon_0 = Y - f_0(T)$ satisfies $\mathbb{E}_P[\varepsilon_0 \mid Z] = 0$ and $\mathbb{E}_P[\varepsilon_0^2] < \infty$, and write $\sigma^2(Z) = \mathbb{E}_P[\varepsilon_0^2 \mid Z]$. (i) Let a be centered, $1 + a \geq 0$, and $\mathbb{E}_P[(1 + |a|)|\varepsilon_0|] < \infty$. For $dP_a = (1 + a)dP$,*

$$\mathbb{E}_{P_a}[\varepsilon_0 \mid Z] = \frac{\mathbb{E}_P[a\varepsilon_0 \mid Z]}{1 + \mathbb{E}_P[a \mid Z]} \quad P_{a,Z}\text{-almost surely}, \tag{8}$$

so P_a preserves the CMR at f_0, if and only if the numerator vanishes $P_{a,Z}$-almost surely. (ii) The preserving scores $\mathcal{T}_0 = \{a \in L_0^2(P) : \mathbb{E}_P[a\varepsilon_0 \mid Z] = 0\}$ form a closed subspace, and

$$L_0^2(P) = \mathcal{T}_0 \oplus \overline{\{\varepsilon_0 h(Z) : h \text{ bounded}\}}. \tag{9}$$

Along a violating direction $a = t\, \varepsilon_0 h(Z)$ with $1 + a \geq 0$, $\mathbb{E}_{P_a}[\varepsilon_0 \mid Z] = t\, h(Z)\sigma^2(Z)$.

Apx. A proves the propositions and bounds the responses of omitted directions.

At the truth, a direction in $\mathcal{T}_0$ has no moment response $\mathcal{D}_{f_0} a = 0$. The full preserving space may contain joint scores, but it depends on the unknown f_0. KS uses an instrument-only subspace that can be chosen without knowing f_0. TAN tests additional joint directions, including ones for which

the response remains nonzero at f_0. To compare these choices as estimators, we need moment responses that can be estimated from data and optimized over.

In practice, we approximate the RKHS using Nyström landmarks. A coefficient vector θ now a structural candidate $f_\theta(t) = \phi(t)^\top\theta$. We test it through the moment vector $g_P(\theta) = b_P - A_P\theta$: $b_P = \mathbb{E}_P[q(Z)Y]$ is the outcome-moment vector, and $A_P = \mathbb{E}_P[q(Z)\phi(T)^\top]$ is the treatment operator mapping θ to its instrument moments. These coordinates restrict the structural class and the moments tested, and give $\|f_\theta\|_{\mathcal{F}} = \|\theta\|_2$. For KS we use kernel test functions h of Z, while full empirical TAN spans all centered reweightings of the observations.

Now we use a shift vector $s = (s_1, \ldots, s_d)$, whose elemets are $a(V) = s(V)^\top u$. To study the exact responses to the score shift for $dP_u = (1 + s^\top u)dP$, we collect them into a matrix D

$$D_j(\theta) = \mathbb{E}_P[s_j(V)q(Z)\psi(T, Y; f_\theta)], \qquad D(\theta) = [D_1(\theta), \ldots, D_d(\theta)] \tag{10}$$

in which each entry is simply weighted by s, obtaining $g + D(\theta)u$ for $g = g_P(\theta)$ (Duchi et al., 2021). After centering, the exact *worst case* moments $g + D(\theta)u$ is simply the farthest point from zero, i.e.,

$$W_\rho(\theta) = \sup_{\|u\|_2 \le \rho} \|g + D(\theta)u\|_2 \ . \tag{11}$$

Unfortunately, it has no general closed-form expression (El Ghaoui & Lebret, 1997). Fortunately, the outer minimization over θ remains convex because g and D are linear.

To obtain efficiently optimizable forms, we consider two lower and upper bounds. For an *upper bound*, we wrap the moments with an Euclidean ball centered at g using a spectral norm or a Frobenius norm

$$U_{\mathrm{op},\rho}(\theta) = \|g\|_2 + \rho\|D(\theta)\|_{\mathrm{op}}, \qquad U_{\mathrm{F},\rho}(\theta) = \|g\|_2 + \rho\|D(\theta)\|_F. \tag{12}$$

For a *lower bound*, we average over all moments, we resort to he root mean squared (RMS) moment error, assuming a uniform relevance of moments. The resulting RMS criterion is

$$R_\tau(\theta) = \left(\|g\|_2^2 + \tau\|D(\theta)\|_F^2\right)^{1/2} \ . \tag{13}$$

By using the same radius τ, we obtain the envelop

$$\underbrace{R_{\rho^2/d}(\theta)}_{\text{RMS}} \le \underbrace{W_\rho(\theta)}_{\text{exact worst case}} \le \underbrace{U_{\mathrm{op},\rho}(\theta)}_{\text{spectral}} \le \underbrace{U_{\mathrm{F},\rho}(\theta)}_{\text{Frobenius}}. \tag{14}$$

In sum, for RMS we optimistically penalize for mean of stacked moments, while Frobenius bounds the responses more pessimistically in relationship to the worst-case direction.

4 Robustness

Here, we study if how this helps us to find structural functions. For this, we first characterize the population displacement of TAN–RMS, then show that the unsquared nominal loss in the upper envelopes can preserve the target over a positive range of radii.

A satisfied CMR guaranteess that the moments disappear at the true function. However, a joint shift can change those moments even at the truth. To this end, we compare IV to least squares fits.

Proposition 2 (The price of joint sensitivity) *Let $\mathcal{S} = L_0^2(P)$ and $0 \le \tau \le 1$, and assume the reference CMR at $f_0 = \phi^\top\theta_0$ with $\|q(Z)\|_2\phi(T)$ and $\|q(Z)\|_2 Y$ square-integrable. For the full score space, define $R_\tau^2 = \|g_P\|_2^2 + \tau\|\mathcal{D}_{f_\theta}\|_{\mathrm{HS}}^2$ using the Hilbert–Schmidt response energy. Then, for every θ,*

$$R_\tau(\theta)^2 = (1-\tau)\|g_P(\theta)\|_2^2 + \tau\,\mathbb{E}_P\big[\|q(Z)\|_2^2(Y - f_\theta(T))^2\big], \tag{15}$$

In particular, $R_\tau(\theta_0)^2 = \tau\mathbb{E}_P[\|q(Z)\|_2^2\varepsilon_0^2]$. Let $A = \mathbb{E}_P[q(Z)\phi(T)^\top]$, $\Sigma_q = \mathbb{E}_P[\|q(Z)\|_2^2\phi(T)\phi(T)^\top]$, and $c_q = \mathbb{E}_P[\|q(Z)\|_2^2\phi(T)\varepsilon_0]$. If $H_\tau = (1-\tau)A^\top A + \tau\Sigma_q$ is invertible, the minimizer is $\theta_\tau^* = \theta_0 + \tau H_\tau^{-1}c_q$.

In Eq. equation 15 we see the tradeoff. We see that at $\tau = 0$ we fit IV moments, but at $\tau = 1$ we fit a regression weighted by instrument-feature strength. Inbetween, the regression term can robustify a

weak IV, but it can also move the estimate toward a confounded model. If c is zero, TAN–RMS does not move the population target, even though the truth remains sensitive to joint shifts. This means, this method is nonparametric varint of a k-class case (Nagar, 1959).

We note that at the truth the IV norm is zero, and moving away increases it massively. However, for a sufficiently small radius, this increase can outweigh the gain from robustness. To study this, we compare the covariances $\Omega(\theta) = \mathrm{Cov}_P(q(Z)(Y - \phi(T)^\top\theta))$ at our model θ and at the turth $\Omega_0 = \Omega(\theta_0)$, obtaining $U_{\mathrm{F},\rho} = \|g_P\|_2 + \rho\sqrt{\mathrm{tr}\,\Omega(\theta)}$ and $U_{\mathrm{op},\rho} = \|g_P\|_2 + \rho\sqrt{\lambda_{\max}(\Omega(\theta))}$. With this, we identify the radius up to which this protection holds.

Proposition 3 (Target-preservation thresholds for the envelopes) *Assume the CMR, integrability, and full TAN setting of Proposition 2, and suppose A has full column rank and $\mathrm{tr}\,\Omega_0 > 0$. Set $\|c\|_{A,*} = \sqrt{c^\top(A^\top A)^{-1}c}$. For unregularized population fitting over $\mathbb{R}^p$:*

(i) θ_0 minimizes $U_{\mathrm{F},\rho}$ if and only if $\rho \le \rho^*_{\mathrm{F}} := \sqrt{\mathrm{tr}\,\Omega_0}/\|c_q\|_{A,*}$.
(ii) If the largest eigenvalue $\eta > 0$ of Ω_0 is simple with unit eigenvector v, set $c_v = \mathbb{E}_P[(v^\top q)^2\phi\varepsilon_0]$. Then θ_0 minimizes $U_{\mathrm{op},\rho}$ if and only if $\rho \le \rho^*_{\mathrm{op}} := \sqrt{\eta}/\|c_v\|_{A,*}$.

Here, a zero denominator means an infinite threshold, and below each threshold the minimizer is unique. The repeated-eigenvalue case is given in Appendix A.

In brief, the thresholds shows the trade off how robustness moves away from the truth and how the IV moments's seeks to stay at the truth.

For example in 1D, considering the $\delta = \theta - \theta_0$, RMS satisfies $R^2_\tau = \delta^2 + \tau(1-\delta)^2$, whereas both envelopes equal $|\delta| + \rho|1-\delta|$ when $T = 1 + U$, $Y = \theta_0 T + U$, and independent Z. That means RMS moves toward the confounded regression with $\tau/(1+\tau)$ as soon as $\tau > 0$. Empirically, we see this effect in Apx. G. In Apx. E we provide conditions under which vanishing moment error yields vanishing structural error on a norm ball.

All proofs are in Apx. A, Appendices C–F provide approximation bounds, RMS identification, and statistical properties of our methods.

5 Related Work

Our work is related to conditional-moment estimation, distributional IV, and distributionally robust optimization. Kernel IV estimates a conditional expectation operator in a first stage and uses it to recover the structural function in a second stage (Singh et al., 2019). Kernel maximum moment loss instead measures residual moments through an instrument kernel (Zhang et al., 2023), while minimax CMR estimators optimize over a class of detecting functions (Dikkala et al., 2020). We build on this moment-based learning signal, using score features to describe perturbations and instrument features to detect violations.

FGEL, KMM, and SMM provide the closest distributional formulations. FGEL estimates a distribution close to the empirical distribution under which the conditional moments hold (Kremer et al., 2022). KMM uses maximum mean discrepancy to extend this construction beyond empirical reweighting (Kremer et al., 2023). SMM uses regularized optimal transport and derives a local expansion involving data derivatives (Kremer & Schölkopf, 2024). We use joint scores or an instrument-feature sieve to specify coupled moment responses, then control their worst-case or average magnitude rather than seeking a distribution that satisfies the moments.

Local divergence DRO already connects distributional perturbations with variance regularization (Duchi et al., 2021), that is shared by TAN–RMS's moment vector. Our pushforward argument and the coupling between treatment and outcome moment sallows us to establish a coupling with robust least-squares (El Ghaoui & Lebret, 1997). Our work also yields a nonparametric sibling of classical anchor regression and k-class estimators (Rothenhäusler et al., 2021; Jakobsen & Peters, 2022) that come with similar minimax guarantees but for linear models. For KS–RMS, using the product instruments $s_j q$ also yields a kernel sieve equation 21.

6 EXPERIMENTS

Having established our theory, we now put it into practice. We study when robustness from joint sensitivity outweigh the confoundedness of standard regression. Specifically, we ask:

- **RQ1** Does sensitivity control improve recovery when instruments are weak, and how does this depend on the radius and sample size?
- **RQ2** Under invalid instruments, training contamination, and test-time attacks, when do joint directions help despite the risk of moving the target?
- **RQ3** Under shifts that preserve or violate the CMR, how do structural error and its dependence on the sensitivity radius change?
- **RQ4** Without structural ground truth, does the method improve over baselines on real IV datasets and under grouped distribution shift?

To measure structural error, we use data-generating processes with known h_0, writing X and h_0 for the treatment T and structural function f_0 in the experiments. The main synthetic and semi-synthetic designs follow the standard IV form

$$\begin{aligned} X &= m(Z) + g_X(U) + \varepsilon_X, \\ Y &= h_0(X) + g_Y(U) + \varepsilon_Y \ , \end{aligned} \tag{16}$$

Here U is a shared, unobserved Gaussian confounder. In our main design, the instrument coordinates are independent uniform draws, the first stage is linear, and h_0 is a coordinatewise absolute-value function. We standardize and scale the instrument signal to target a stated first-stage R^2. Most remaining treatment variation comes from the shared confounder. Independent treatment noise is small. In the semi-synthetic design, image features supply the instruments while the structural relationship remains known.

In the default design, we estimate from $4\,000$ training observations with four treatment dimensions and four instruments. Our kernel representations use 100 treatment and 200 instrument Nyström features, with ranks capped for smaller samples. Our criteria use a fixed HC0 moment transform from a same-sample pilot and, in synthetic and semi-synthetic comparisons, share a penalty chosen by an empirical-Bayes-inspired (EB) score on TAN–RMS (Appendix H). Competitors use five-fold cross-validation, which Appendix J also applies to our criteria. The main displays compare TAN–RMS, TAN–Frobenius, and KS–RMS. The first isolates joint average response, the second controls an upper bound, and the third tests preserving scores under the RMS objective (equation 21). Appendix J reports TAN–Spectral and KS–Frobenius at the fixed radius. The spectral radius path was not run. Baselines are MMR-IV (Zhang et al., 2023), Sieve 2SLS (Newey & Powell, 2003), KIV (Singh et al., 2019), FGEL (Kremer et al., 2022), and SMM (Kremer & Schölkopf, 2024). Appendix H gives the implemented objectives. We set the RMS coordinate variance to $\tau = \kappa/n$ and the spectral and Frobenius radii to $\rho = \sqrt{\kappa/n}$, with $\kappa = 3$ unless it is the experimental axis. The common κ varies sensitivity but does not match the radii in equation 14 (Appendix C). To assess recovery when the truth is known, we report *structural mean-squared error*

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\hat{f}(x_i) - h_0(x_i))^2$$

on $1\,000$ independent test observations, or $2\,000$ in the joint-shift experiments. On real data, where h_0 is unobserved, we instead report projected residual error on the held-out 20% of observations, as defined below. Figures report means and 95% intervals across ten seeds. The real-data table reports standard errors.

6.1 IDENTIFICATION, RADIUS, AND SAMPLE SIZE

Weak instruments make it harder to distinguish structural functions, while a larger sensitivity weight can stabilize or displace the fit. We vary first-stage strength, sensitivity scale κ, and sample size to see both effects (Figs. 1, 3, and 2), using $\kappa = 3$ outside the sensitivity sweep. While all methods worsen with weaker first stages, ORION almost always retains the lowest error throughout. When we increase the sample size all methods benefit, but ORION remains in the lead (Fig. 2, left). With 250 or 500 observations, KS–RMS has the lowest error and TAN matches the best competitor; from $n = 1\,000$ on, TAN leads, and at $n = 4\,000$ TAN–RMS and TAN–Frobenius reach 0.70 and

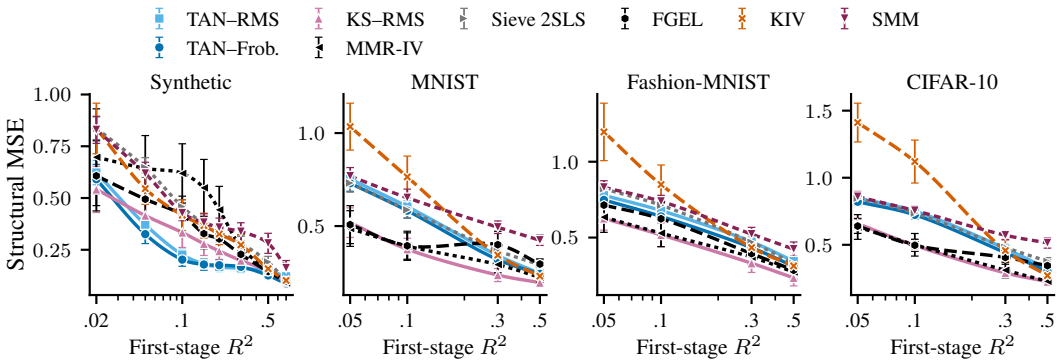


Figure 1: Our methods are robust against weak instruments. We show structural MSE against first-stage strength with synthetic instruments and with image-feature instruments (MNIST, Fashion-MNIST, CIFAR-10).

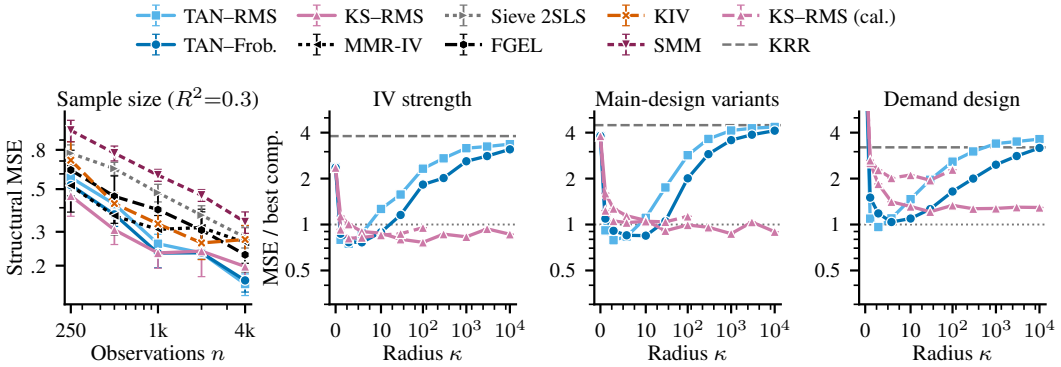


Figure 2: Sampling noise impacts our methods the least, and KS is resilient against misspecified radii. We report structural MSE against training sample size (R^2=0.3, left), and structural MSE relative to the best competitor along the radius path up to $\kappa = 10^4$ for three benchmark groups (right). The dashed line is indicates the baseline kernel ridge regression.

0.74 of the best competitor's error. Along the extended radius paths (Fig. 2, right), TAN keeps a low error over small radii before rising toward the error of regression without instruments (KRR), as Proposition 2 predicts. TAN–Frobenius stays flat over a wider range of radii than TAN–RMS, consistent with the target-preservation thresholds of Proposition 3. KS–RMS changes little with κ, reflecting its much smaller response energy at the standard-error scale (Appendix G).

6.2 Semi-Synthetic Instruments

To study semi-synthetic data, we replace synthetic instruments by standardized PCA components from MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), and CIFAR-10 (Krizhevsky, 2009), retaining the structural relationship so that the causal targets remain known. In Fig. 1, ORION, led by KS, recovers the truth best on all three datasets except at the weakest strength, where MMR-IV is lowest on MNIST and CIFAR-10.

6.3 Violations, Contamination, and Attacks

We study exclusion violation and endogeneity in the invalid-instrument design of Qu & Kwon (2024), training contamination in the design of Kremer & Schölkopf (2024), and test-time attacks in our main design (Appendix I). First, a *direct linear effect* $Z \to Y$ of varying strength violates exclusion. Second, a direct effect $U \to Z$ of varying strength makes the instrument endogenous. Third, we replace a fraction of training treatments by random values from their observed range and

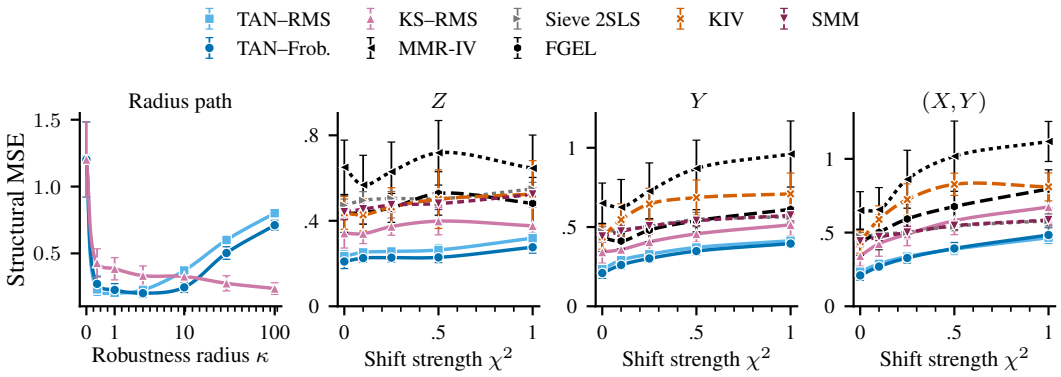


Figure 3: Our method has the lowest error and is robust in any case. We report the structural MSE against robustness scale and against shift strength when training data are reweighted toward Z, Y, or (X, Y) for R^2=0.1

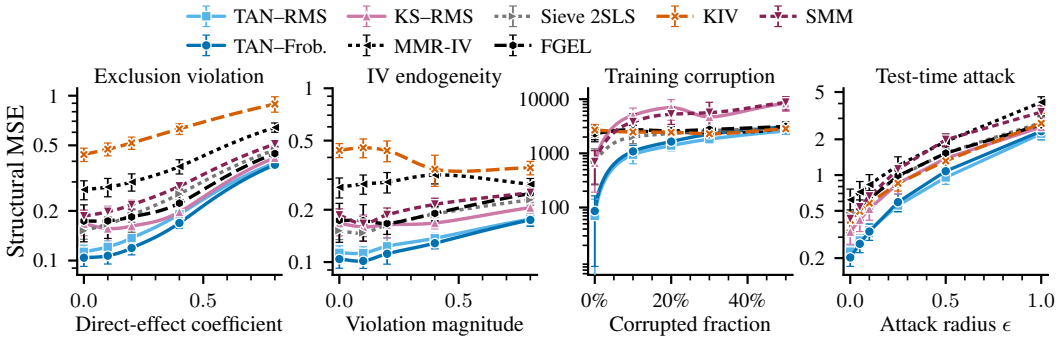


Figure 4: Our method has the lowest error and is robust in any case. We report the structural MSE under exclusion violations, instrument endogeneity, corrupted training treatments, and adversarial test-treatment perturbations.

keep test data clean. Finally, we perturb each test treatment in the gradient-sign direction of the fitted model's squared structural error, measuring size in treatment standard deviations; since the fit and the known truth are both evaluated at the perturbed treatment, this is an evaluation attack. In Fig. 4, all methods degrade with increasing exclusion violation and endogeneity, but ORION retains lower mean error. Under training contamination, errors rise sharply, and TAN–RMS stays lowest throughout. Under test-time attacks, TAN retains the lowest error, while MMR-IV and SMM degrade fastest.

6.4 JOINT TRAINING–TARGET SHIFTS

We next ask what happens when the training sample differs from the population whose structural function we want to recover. We select $n = 4\,000$ observations from a pool of $10n$ target draws using weights based on Z, Y, or (X, Y) (Appendix H), then evaluate on the untilted population. The horizontal axis gives the nominal Pearson divergence of the pool weights. Selection based on Z preserves the CMR. Selection based on Y can change it even at the true function (Proposition 1). Thus fitting the training moments and recovering the target function can become different goals. Joint directions can detect more of this change, but sensitivity control alone cannot restore the target moments.

Weights based on Z keep the CMR, and errors grow by at most $1.4\times$ at $\chi^2 = 1$ (Fig. 3). Under outcome-dependent selection, errors grow by up to $2.3\times$, with TAN growing fastest. This is consistent with the risk of target displacement established for TAN–RMS in Proposition 2, though that result does not describe the fitted Frobenius or spectral criteria. Under weak identification ($R^2 = 0.1$),

Table 1: Our methods win on most real-world data sets, and are highly competitive in each case. We report the means ± standard errors of the projected residual error (PRE) (↓) across ten seeds, with *cross-validating* paramters for all methods.

Dataset	Split	TAN–RMS	TAN–Frob.	KS–RMS	MMR-IV	FGEL	KIV	SMM
Airline	Random	**0.0313**$_{\pm 0.0018}$	0.0316$_{\pm 0.0017}$	0.0344$_{\pm 0.0019}$	0.0601$_{\pm 0.0128}$	0.0353$_{\pm 0.0019}$	0.0630$_{\pm 0.0130}$	0.0377$_{\pm 0.0020}$
Automobile		**0.128**$_{\pm 0.0106}$	0.135$_{\pm 0.0118}$	0.152$_{\pm 0.0141}$	0.195$_{\pm 0.0165}$	0.173$_{\pm 0.0162}$	0.164$_{\pm 0.0081}$	0.130$_{\pm 0.0074}$
Fulton Fish		**0.737**$_{\pm 0.130}$	0.761$_{\pm 0.146}$	0.832$_{\pm 0.161}$	1.12$_{\pm 0.218}$	0.984$_{\pm 0.170}$	1.05$_{\pm 0.177}$	1.01$_{\pm 0.158}$
JTPA		0.0572$_{\pm 0.0036}$	0.0577$_{\pm 0.0037}$	**0.0572**$_{\pm 0.0039}$	0.0608$_{\pm 0.0042}$	0.0577$_{\pm 0.0041}$	0.0616$_{\pm 0.0046}$	0.0587$_{\pm 0.0036}$
401(k)		193.99$_{\pm 38.81}$	198.82$_{\pm 40.45}$	199.70$_{\pm 40.77}$	204.43$_{\pm 48.77}$	208.46$_{\pm 46.66}$	203.51$_{\pm 42.63}$	**191.27**$_{\pm 40.03}$
Airline	Shift	**0.0342**$_{\pm 0.0020}$	0.0346$_{\pm 0.0020}$	0.0373$_{\pm 0.0018}$	0.0543$_{\pm 0.0024}$	0.0389$_{\pm 0.0017}$	0.0529$_{\pm 0.0013}$	0.0396$_{\pm 0.0019}$
Automobile		**0.128**$_{\pm 0.0154}$	0.144$_{\pm 0.0197}$	0.158$_{\pm 0.0177}$	0.234$_{\pm 0.0229}$	0.192$_{\pm 0.0211}$	0.207$_{\pm 0.0112}$	0.136$_{\pm 0.0144}$
Fulton Fish		0.809$_{\pm 0.248}$	**0.784**$_{\pm 0.236}$	0.922$_{\pm 0.300}$	1.50$_{\pm 0.416}$	1.23$_{\pm 0.327}$	1.31$_{\pm 0.350}$	1.05$_{\pm 0.281}$
JTPA		0.0592$_{\pm 0.0050}$	0.0592$_{\pm 0.0050}$	0.0589$_{\pm 0.0051}$	0.0574$_{\pm 0.0048}$	0.0586$_{\pm 0.0050}$	0.0592$_{\pm 0.0045}$	**0.0568**$_{\pm 0.0046}$
401(k)		429.08$_{\pm 137.23}$	435.89$_{\pm 137.43}$	431.49$_{\pm 136.45}$	443.52$_{\pm 136.92}$	440.97$_{\pm 135.87}$	434.00$_{\pm 120.83}$	**365.38**$_{\pm 111.20}$

TAN retains the lowest error across all shift strengths and all three directions. TAN–Frobenius reaches 0.70 and 0.83 of the best competitor's error at $\chi^2 = 1$ for shifts toward Y and (X, Y). For shifts toward (X, Y), KS–RMS starts below the competitors but loses its lead by $\chi^2 \approx 0.25$. The advantage narrows as the shift and the first stage strengthen: at $R^2 = 0.3$ and $\chi^2 = 1$, TAN and Sieve 2SLS have similar errors. The recovery advantage does not establish ambiguity-set coverage, which depends on the reverse divergence (Appendix H).

6.5 Real-World Datasets

Finally, we finally on real-world datasets. *Airline* uses U.S. domestic market data from the DB1B survey (BTS, 2025), with log fare as treatment, log passengers as outcome, and leave-market-out fares as instruments (grouped splits group on markets). *Automobile* uses the demand data of Berry et al. (1995) with the excluded demand instruments (Conlon & Gortmaker, 2020) (grouped splits split on markets). *Fulton Fish* uses the daily fish-market data of Graddy (1995), with log price as treatment, log quantity as outcome, and sea weather (wind speed and wave height) as instruments (grouped splits hold out blocks of 20 trading days). *JTPA* uses the randomized training offer as instrument for participation, with earnings as outcome (Bloom et al., 1997) (grouped splits split on one sex). *401(k)* uses plan eligibility as instrument for participation, with net financial assets as outcome (Abadie, 2003) (grouped splits split on age buckets).

Since the truth is unknown, we measure residual variation predictable from instruments (Dikkala et al., 2020) on held-out data. We smooth residuals $r_i = y_i - \hat{f}(x_i)$ with $S = K_Z(K_Z + n_{\text{test}}\gamma I)^{-1}$, where K_Z is the instrument Gram matrix with bandwidth chosen from training instruments and $\gamma = 10^{-3}$, and report projected residual error (PRE) as $\frac{1}{n_{\text{test}}}\|Sr\|_2^2$. This diagnostic retains residual noise and does not measure structural error. We use both random and dataset-specific grouped splits to assess performance across groups. TAN–RMS obtains the lowest mean held-out error for Airline and Automobile under both splits, and a TAN criterion leads on both Fulton Fish splits, although its 97 daily observations leave large standard errors. KS–RMS leads on the random JTPA split, where its displayed mean equals that of TAN–RMS, and SMM leads on the random 401(k) split and on both grouped JTPA and 401(k) splits.

7 Conclusion

We studied which distributional changes an IV estimator should control. The pushforward retains the coupled changes in treatment operators and outcome moments. The norm sandwich gives simpler ways to control or average their responses. Instrument-only scores (KS) control a preserving subspace. Joint scores (TAN) control broader responses, but can also affect the structural target. The RMS displacement formula connects stabilization to weighted regression and its possible confounding bias. Under full-rank identification, the upper envelopes retain the population target below explicit sensitivity thresholds. Empirically, the proposed criteria improve recovery in the studied weak-instrument and invalid-instrument settings and retain an advantage under moderate joint shifts, while real-data results concern held-out moments. Our guarantees cover absolutely continuous shifts and independently fixed features. Other shifts and the learned features and metric remain open.

AI USE STATEMENT

In this work, we used generative AI tools for proof-reading, feedback on the presentational aspects, and to get feedback for contributions and correctness. We used it to seek references we might have missed, from which we drew connections to classical work in different research disciplines. We have neither used generative AI tools for formulating the research questions or for answering it. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 6 describes the evaluation in detail, Appendix H specifies tuning and parameters, and Appendix J reports additional results. All assumptions are clearly stated, and Appendix A provides our proofs. Appendices C–F give the approximation bounds, RMS identification, and statistical conditions.

REFERENCES

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995. doi: 10.2307/2171802.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Howard S. Bloom, Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Johannes M. Bos, and Robert G. Wallace. The benefits and costs of JTPA Title II-A programs: Key findings from the National JTPA Study. *The Journal of Human Resources*, 32(3):549–576, 1997. doi: 10.2307/146183.

BTS. Bureau of Transportation Statistics. U.S. Airline Traffic Data and Demand Indicators. U.S. Department of Transportation. https://www.bts.gov, 2025. Accessed: 2026-09-12.

Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In *Handbook of Econometrics*, volume 6B, pp. 5633–5751. Elsevier, 2007.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(87)90015-7.

Christopher Conlon and Jeff Gortmaker. Best practices for differentiated products demand estimation with PyBLP. *The RAND Journal of Economics*, 51(4):1108–1161, 2020.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12248–12262. Curran Associates, Inc., 2020.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Pavlos S. Efraimidis and Paul G. Spirakis. Weighted random sampling with a reservoir. *Information Processing Letters*, 97(5):181–185, 2006.

Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, 1997.

Kathryn Graddy. Testing for imperfect competition at the Fulton fish market. *The RAND Journal of Economics*, 26(1):75–92, 1995. doi: 10.2307/2555636.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pp. 1414–1423. PMLR, 2017.

Martin Emil Jakobsen and Jonas Peters. Distributional robustness of K-class estimators and the PULSE. *The Econometrics Journal*, 25(2):404–432, 2022.

Heiner Kremer and Bernhard Schölkopf. Geometry-aware instrumental variable regression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25560–25582. PMLR, 2024. URL https://proceedings.mlr.press/v235/kremer24a.html.

Heiner Kremer, Jia-Jie Zhu, Krikamol Muandet, and Bernhard Schölkopf. Functional generalized empirical likelihood estimation for conditional moment restrictions. In *International Conference on Machine Learning*, pp. 11665–11682. PMLR, 2022.

Heiner Kremer, Yassine Nemmour, Bernhard Schölkopf, and Jia-Jie Zhu. Estimation beyond data reweighting: Kernel method of moments. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17745–17783. PMLR, 23–29 Jul 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pp. 148–188. Cambridge University Press, 1989.

Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*, pp. 41–50. PMLR, 2020.

A. L. Nagar. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, 27(4):575, October 1959. ISSN 0012-9682. doi: 10.2307/1909352. URL http://dx.doi.org/10.2307/1909352.

Whitney K. Newey. Efficient estimation of models with conditional moment restrictions. In *Econometrics*, volume 11 of *Handbook of Statistics*, pp. 419–454. Elsevier, 1993. doi: https://doi.org/10.1016/S0169-7161(05)80051-3. URL https://www.sciencedirect.com/science/article/pii/S0169716105800513.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Whitney K. Newey and Richard J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004. ISSN 00129682, 14680262.

Zhaonan Qu and Yongchan Kwon. Distributionally robust instrumental variables estimation. *arXiv preprint arXiv:2410.15634*, 2024.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

Wenqi Shi and Wenkai Xu. Learning nonlinear causal effects via kernel anchor regression. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1942–1952, 2023.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, 2001.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. doi: 10.48550/arXiv.1708.07747.

Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1):20220073, 2023.

A Proofs of the main results

Proof of the omitted-direction bounds. Write $\mathcal{W}_{\mathcal{S}}(f;\rho) = \sup_{a\in\mathcal{S},\,\|a\|_{L^2(P)}\le\rho}\|\mu_P(f) + \mathcal{D}_f a\|_{\mathcal{H}}$. We show $0 \le \mathcal{W}_{L^2_0(P)}(f;\rho) - \mathcal{W}_{\mathcal{S}}(f;\rho) \le \rho\|\mathcal{D}_f(I-\Pi_{\mathcal{S}})\|_{\mathrm{op}}$. The lower bound follows from set inclusion. For the upper bound, decompose $a = \Pi_{\mathcal{S}}a + (I-\Pi_{\mathcal{S}})a$, use $\|\Pi_{\mathcal{S}}a\| \le \|a\|$, and apply the triangle inequality. Taking the supremum over the full ball proves the upper bound, since $\|\mathcal{D}_f(I-\Pi_{\mathcal{S}})a\| \le \rho\|\mathcal{D}_f(I-\Pi_{\mathcal{S}})\|_{\mathrm{op}}$. For RMS, extend an orthonormal basis of $\mathcal{S}$ to one of $L^2_0(P)$. The squared Hilbert–Schmidt norm is the sum of squared responses over this basis, so the full response energy splits into $\|\mathcal{D}_f\Pi_{\mathcal{S}}\|^2_{\mathrm{HS}} + \|\mathcal{D}_f(I-\Pi_{\mathcal{S}})\|^2_{\mathrm{HS}}$.

Proof of Proposition 1. For (i), let $w = 1 + a \ge 0$. Conditional change of measure gives $\mathbb{E}_{P_a}[\varepsilon_0 \mid Z] = \mathbb{E}_P[w\varepsilon_0 \mid Z]/\mathbb{E}_P[w \mid Z]$ where the denominator is positive. Its zero set has zero $P_{a,Z}$ mass. The reference CMR reduces the numerator to $\mathbb{E}_P[a\varepsilon_0 \mid Z]$, which proves both the identity and the equivalence.

For (ii), the map $a \mapsto \mathbb{E}_P[a\varepsilon_0 \mid Z]$ is continuous from $L^2(P)$ to $L^1(P)$ by Cauchy–Schwarz, so $\mathcal{T}_0$ is closed. For bounded h, the score $\varepsilon_0 h(Z)$ is centered because $\mathbb{E}_P[\varepsilon_0 h(Z)] = \mathbb{E}_P[h(Z)\mathbb{E}_P[\varepsilon_0 \mid Z]] = 0$, and

$$\langle a, \varepsilon_0 h(Z)\rangle_{L^2(P)} = \mathbb{E}_P\big[h(Z)\,\mathbb{E}_P[a\varepsilon_0 \mid Z]\big].$$

This vanishes for every bounded h if and only if $\mathbb{E}_P[a\varepsilon_0 \mid Z] = 0$. Hence $\mathcal{T}_0$ is the orthogonal complement of the closed span of the violating scores, which proves equation 9. Finally, for $a = t\varepsilon_0 h(Z)$, we have $\mathbb{E}_P[a\varepsilon_0 \mid Z] = t\,h(Z)\sigma^2(Z)$ and $\mathbb{E}_P[a \mid Z] = t\,h(Z)\mathbb{E}_P[\varepsilon_0 \mid Z] = 0$, and (i) gives the last claim.

Representing a prescribed conditional-moment change $v(Z)$ by $a = \varepsilon_0 v(Z)/\sigma^2(Z)$ requires $v = 0$ where $\sigma^2 = 0$ and $\mathbb{E}_P[v^2/\sigma^2] < \infty$, with the ratio set to zero on that set. These conditions put the score in the closed complement in equation 9; the corresponding finite reweighting still requires $1 + a \ge 0$. In particular, for two-sided unbounded errors, a nonzero linear score $t\varepsilon_0 h(Z)$ can fail positivity at every nonzero t even though it is a valid tangent direction.

Proof of Proposition 2. Complete an orthonormal basis (s_j) of $L^2_0(P)$ by the constant function to one of $L^2(P)$. Parseval's identity, applied to each coordinate of $\xi = q(Z)\psi(T,Y;f_\theta)$, gives $\sum_j \|\mathbb{E}_P[s_j\xi]\|_2^2 = \mathbb{E}_P\|\xi\|_2^2 - \|\mathbb{E}_P\xi\|_2^2$. Hence $R_\tau^2 = \|g_P\|_2^2 + \tau(\mathbb{E}_P[\|q(Z)\|_2^2\psi^2] - \|g_P\|_2^2)$, which is equation 15. At θ_0, the reference CMR gives $g_P(\theta_0) = 0$ and $\psi = \varepsilon_0$, so $R_\tau(\theta_0)^2 = \tau\,\mathbb{E}_P[\|q(Z)\|_2^2\varepsilon_0^2] = \tau\,\mathbb{E}_P[\|q(Z)\|_2^2\sigma^2(Z)]$. For $0 \le \tau \le 1$, equation 15 is a convex quadratic in θ with normal equations $H_\tau\theta = (1-\tau)A^\top b_P + \tau\,\mathbb{E}_P[\|q(Z)\|_2^2\phi(T)Y]$. Substituting $b_P = A\theta_0$ and $\mathbb{E}_P[\|q(Z)\|_2^2\phi(T)Y] = \Sigma_q\theta_0 + c_q$ gives $H_\tau\theta = H_\tau\theta_0 + \tau c_q$, which proves the displacement formula.

Proof of Proposition 3. Let $h = \theta - \theta_0$ and $\xi_\theta = q(Y - \phi^\top\theta)$. The CMR gives $g_P(\theta) = -Ah$. The centered random vector $\xi_\theta - \mathbb{E}_P\xi_\theta$ is affine in θ, so its L^2 norm and its induced spectral response norm are convex. At θ_0, differentiating the covariance gives

$$\Omega'(\theta_0)[h] = -2\mathbb{E}_P[qq^\top\varepsilon_0\phi^\top h],$$

because $g_P(\theta_0) = 0$. Writing $v_0 = \sqrt{\operatorname{tr}\Omega_0}$, the gradient of the Frobenius response norm is therefore $-c_q/v_0$. The subdifferential of the nominal norm at θ_0 is $\{A^\top u : \|u\|_2 \le 1\}$. Convex optimality is consequently equivalent to $A^\top u = \rho c_q/v_0$ for some $\|u\|_2 \le 1$. For full-column-rank A, the minimum norm of a solution of $A^\top u = z$ is $\|z\|_{A,*}$, proving (i).

For spectral, write the response norm as

$$\sqrt{\lambda_{\max}(\Omega(\theta))} = \max_{\|v\|_2=1} \sqrt{\operatorname{Var}_P(v^\top \xi_\theta)}.$$

Every active v at θ_0 lies in the top eigenspace and has positive variance η. The corresponding gradient is $-\mathbb{E}_P[(v^\top q)^2 \phi\varepsilon_0]/\sqrt{\eta}$. For a repeated top eigenvalue, let $\mathcal{M}_0 = \{M \succeq 0 : \operatorname{tr} M = 1,\ \Omega_0 M = \eta M\}$ and $c_M = \mathbb{E}_P[(q^\top M q)\phi\varepsilon_0]$. The subdifferential of this maximum is the convex hull of these gradients, namely $\{-c_M/\sqrt{\eta} : M \in \mathcal{M}_0\}$. Thus optimality is equivalent to $A^\top u = \rho c_M/\sqrt{\eta}$ for some $M \in \mathcal{M}_0$ and $\|u\|_2 \le 1$. The set $\mathcal{M}_0$ is compact, so the threshold is $\rho^*_{\mathrm{op}} = \sqrt{\eta}/\min_{M\in\mathcal{M}_0} \|c_M\|_{A,*}$, including the simple-eigenvalue case.

For either envelope, choose the subgradient above with minimum dual norm, and call its negative z. The supporting-hyperplane inequality gives, for every h,

$$U(\theta_0 + h) - U(\theta_0) \ge \|Ah\|_2 - \rho z^\top h \ge (1 - \rho\|z\|_{A,*})\|Ah\|_2.$$

Below the threshold this is strictly positive for $h \neq 0$, proving uniqueness. A zero minimum dual norm gives the same strict inequality for every finite radius. No conclusion about uniqueness is asserted at the threshold.

With a ridge term $\lambda\|\theta\|_2^2/2$, the exact Frobenius condition becomes

$$\|\rho c_q/v_0 - \lambda\theta_0\|_{A,*} \le 1,$$

and the spectral condition becomes

$$\min_{M\in\mathcal{M}_0} \|\rho c_M/\sqrt{\eta} - \lambda\theta_0\|_{A,*} \le 1.$$

These need not describe an interval starting at zero. Without full column rank, the corresponding existence conditions $A^\top u = \rho c_q/v_0 - \lambda\theta_0$ or $A^\top u = \rho c_M/\sqrt{\eta} - \lambda\theta_0$, with $\|u\|_2 \le 1$, remain valid, but a positive target-preservation radius is not automatic.

Proof of Proposition 4. Since $\tau > 0$, the squared norm of $K_\tau v$ is a sum of positively weighted squared block norms. It vanishes exactly when each block vanishes. Under target preservation all moment blocks vanish at θ_0, so $R_\tau(\theta) = \|K_\tau(\theta - \theta_0)\|$; its zero set is $\theta_0 + \ker K_\tau$. Write $H = K_\tau^\top K_\tau$ and $\Sigma_w = \mathbb{E}_{P_w}[\phi(T)\phi(T)^\top]$. The inequality requires $\ker H \subseteq \ker \Sigma_w$. Conversely, on the orthogonal complement of $\ker H$, the matrix H is positive definite. Its smallest eigenvalue and the largest eigenvalue of Σ_w give a finite comparison constant there. Both quadratic forms vanish on $\ker H$, so the comparison extends to the whole space.

Proof of Proposition 5. Let $F_\tau(\theta)$ stack the nominal and scaled score moments, so that $R_\tau(\theta) = \|F_\tau(\theta)\|$ and $F_\tau(\theta) - F_\tau(\theta_0) = -K_\tau(\theta - \theta_0)$. The triangle inequality gives $\|K_\tau(\widehat{\theta} - \theta_0)\| \le R_\tau(\widehat{\theta}) + b_\tau$. Approximate empirical optimality and nonnegativity of the fitted ridge penalty imply

$$R_\tau(\widehat{\theta}) \le b_\tau + e_\tau(\theta_0) + e_\tau(\widehat{\theta}) + \lambda\|\theta_0\|^2/2 + \zeta.$$

Combining the two inequalities with equation 17 proves the claim. For example, if $F_\tau(\theta) = J_\tau[\theta^\top, 1]^\top$ and $\|\widehat{J}_\tau - J_\tau\|_{\mathrm{op}} \le \epsilon$, the reverse triangle inequality supplies $e_\tau(\theta) = \epsilon\sqrt{1 + \|\theta\|^2}$. This is a conditional deterministic event, not a rate for adaptive features.

B FINITE-FEATURE IDENTIFICATION AND STRUCTURAL-ERROR TRANSFER

We consider structural error as the squared difference $\mathbb{E}_{P_w}[(\widehat{f}(T) - f_0(T))^2]$ between the fitted structural function and the truth under a target distribution P_w. Transferring moment error to structural error through an identification condition is standard in IV inverse problems (Carrasco et al., 2007;

Dikkala et al., 2020); our question is how the chosen score responses change that condition and the target. For $d \geq 1$, let $A = \mathbb{E}_P[q\phi^\top]$, $A_j = \mathbb{E}_P[s_j q\phi^\top]$, and $\tau > 0$. The RMS response defines the linear map $K_\tau v = (Av, \sqrt{\tau}A_1 v, \ldots, \sqrt{\tau}A_d v)$, whose norm measures the moment response to changing the structural coefficients.

Proposition 4 (Identification by score moments) *The null space is* $\ker K_\tau = \ker A \cap \bigcap_{j=1}^d \ker A_j$. *For instrument-only scores preserving the CMR at a representable* f_0*, the unregularized population RMS criterion identifies its coefficients uniquely if and only if this intersection is* $\{0\}$. *For a target distribution* P_w *with finite feature second moments, a finite constant* $c_{w,\tau}$ *satisfying*

$$\mathbb{E}_{P_w}[(\phi(T)^\top v)^2] \leq c_{w,\tau}\|K_\tau v\|_2^2 \quad \textit{for all } v \tag{17}$$

exists exactly when every vector in $\ker K_\tau$ *has zero target second moment.*

Extra score moments can improve a finite dictionary's identification or conditioning, but instrument-only scores cannot remove population nonidentification: if $\mathbb{E}_P[h(T) \mid Z] = 0$, all integrable instrument-score products also have zero moment with h. Joint scores can remove additional null directions, but do not generally preserve the target; Proposition 2 characterizes their displacement. A large $c_{w,\tau}$ expresses weak identification (Dikkala et al., 2020). To expose the cost of sensitivity control, let R_τ be the population RMS criterion and let $b_\tau = R_\tau(\theta_0)$ for a representable truth $f_0 = \phi^\top\theta_0$.

Proposition 5 (Structural error and the cost of sensitivity) *Suppose equation 17 holds and, on the coefficient set under consideration, containing* θ_0 *and* $\widehat{\theta}$, $|\widehat{R}_\tau(\theta) - R_\tau(\theta)| \leq e_\tau(\theta)$. *If* $\widehat{\theta}$ *minimizes* $\widehat{R}_\tau(\theta) + \lambda\|\theta\|_2^2/2$ *to objective accuracy* $\zeta \geq 0$*, then*

$$\mathbb{E}_{P_w}[(f_{\widehat{\theta}}(T) - f_0(T))^2] \leq c_{w,\tau}\big(2b_\tau + e_\tau(\widehat{\theta}) + e_\tau(\theta_0) + \lambda\|\theta_0\|_2^2/2 + \zeta\big)^2. \tag{18}$$

Under the reference CMR and instrument-only scores, $b_\tau = 0$, so sensitivity control does not itself displace the population truth; for TAN, $b_\tau > 0$ arises from the violating directions of Proposition 1. This deterministic bound separates sensitivity at the truth from moment-estimation error, regularization, optimization, and identification; Appendix F gives a sampling bound for features fixed independently of the estimation sample, unlike our same-sample features.

C Norm sandwich and approximation bounds

Fix a candidate θ, a radius $\rho \geq 0$, and a score dimension $d \geq 1$. Write $g = g_P(\theta)$, $D = D(\theta)$, and $a = \|g\|_2$, $b = \rho\|D\|_{\mathrm{op}}$. Uniform scores on the radius-ρ sphere have mean zero and covariance $\rho^2 I_d/d$. Expanding the squared moment norm therefore gives

$$\mathbb{E}\|g + Du\|^2 = a^2 + \frac{\rho^2}{d}\|D\|_F^2.$$

Every score in this average is feasible for W_ρ, so its square root $R_{\rho^2/d}$ is a lower bound. The triangle inequality gives $W_\rho \leq a + b = U_{\mathrm{op},\rho}$, and $\|D\|_{\mathrm{op}} \leq \|D\|_F$ gives the Frobenius upper bound. Together these prove equation 14.

To understand the approximation gap, choose a unit right singular vector v for the largest singular value of D. Both ρv and $-\rho v$ are feasible, and the average of their squared moment norms is $a^2 + b^2$. Consequently,

$$\sqrt{a^2 + b^2} \leq W_\rho \leq U_{\mathrm{op},\rho} = a + b \leq \sqrt{2}\, W_\rho. \tag{19}$$

The spectral upper bound loses the alignment between the nominal moment and the longest response axis, but this costs at most a factor $\sqrt{2}$. For $D \neq 0$, define its stable rank $r_{\mathrm{eff}} = \|D\|_F^2/\|D\|_{\mathrm{op}}^2$. Then Cauchy–Schwarz gives

$$U_{\mathrm{F},\rho} = a + \sqrt{r_{\mathrm{eff}}}\, b \leq \sqrt{1 + r_{\mathrm{eff}}}\sqrt{a^2 + b^2} \leq \sqrt{1 + r_{\mathrm{eff}}}\, W_\rho.$$

When $D = 0$, all criteria equal $\|g\|$. Frobenius therefore agrees with spectral for a rank-one response, but can be more conservative when several response axes have comparable lengths.

The spherical RMS lower bound need not be close to the worst case: with $g = 0$ and rank-one D, $W_\rho = \sqrt{d}\, R_{\rho^2/d}$. More generally, another application of Cauchy–Schwarz gives $U_{\mathrm{F},\rho} \le \sqrt{d+1}\, R_{\rho^2/d}$. The averaging criterion can therefore underweight a small number of highly sensitive directions, which explains why its computational simplicity does not come with the spectral envelope's dimension-free approximation.

At the experimental covariance scale $\tau = \rho^2$, the same norms instead give

$$\frac{R_{\rho^2}}{\sqrt{d}} \le W_\rho \le U_{\mathrm{op},\rho} \le U_{\mathrm{F},\rho} \le \sqrt{2}\, R_{\rho^2}. \tag{20}$$

The first inequality follows from $R_{\rho^2}/\sqrt{d} \le R_{\rho^2/d}$; the last follows from $a + \rho\|D\|_F \le \sqrt{2}\sqrt{a^2 + \rho^2\|D\|_F^2}$. In particular, R_{ρ^2} itself may lie on either side of W_ρ. These are pointwise comparisons of moment losses, before adding ridge. The ordering in equation 14 also holds after adding the same nonnegative penalty, but does not order structural prediction errors.

D RMS IDENTITIES AND THE k-CLASS SPECIALIZATION

The score construction changes which residual patterns the estimator penalizes. Its finite-coordinate form makes this change explicit and allows RMS to be computed without repeatedly optimizing over scores. Throughout this section, the feature maps and any moment metric are fixed. A metric is absorbed into q by applying the same linear transform to nominal and score moments.

For instrument-only scores, stack $\widetilde{q}(z) = (q(z)^\top, \sqrt{\tau}s_1(z)q(z)^\top, \dots, \sqrt{\tau}s_d(z)q(z)^\top)^\top$. Then the squared norm of $\mathbb{E}[\widetilde{q}(Z)(Y - f_\theta(T))]$ is exactly $R_\tau(\theta)^2$. Consequently, the induced instrument kernel is

$$\widetilde{k}(z, z') = q(z)^\top q(z')\big(1 + \tau s(z)^\top s(z')\big). \tag{21}$$

This follows by taking the inner product of the stacked features. It also holds for their empirical averages. Thus KS-RMS is a kernel moment estimator whose additional instruments are specified by the chosen score responses. Under a valid CMR, these products retain zero moments at the truth whenever they are integrable. Their value is in distinguishing alternatives that the nominal dictionary does not distinguish, rather than changing the population CMR.

For full TAN, let $q_i = q(z_i)$ and $r_i = y_i - f_\theta(t_i)$. An empirically orthonormal basis of centered scores satisfies $\sum_{j=1}^{n-1} s_j(i)s_j(l) = n\mathbf{1}\{i = l\} - 1$. Substitution into $\widehat{D}_j = n^{-1}\sum_i s_j(i)q_i r_i$ gives

$$\widehat{D}\widehat{D}^\top = \frac{1}{n}\sum_i q_i q_i^\top r_i^2 - \widehat{g}\widehat{g}^\top. \tag{22}$$

This covariance identity describes signed tangent scores. A finite reweighting must additionally have nonnegative weights; the full score ball need not satisfy this at the experimental radii. Taking traces gives the empirical TAN criterion

$$\widehat{R}_\tau(\theta)^2 = (1 - \tau)\|\widehat{g}(\theta)\|_2^2 + \frac{\tau}{n}\sum_i \|q_i\|_2^2 r_i^2. \tag{23}$$

Although its mixture interpretation uses $0 \le \tau \le 1$, the original sum of squared affine responses is nonnegative and defines a convex norm for every $\tau \ge 0$. At the population level, the identity also holds with the full instrument RKHS, replacing $\|q(Z)\|_2^2$ by $k(Z, Z)$ under the corresponding square-integrability assumptions. For $k(z, z) = 1$, it combines squared MMR and mean squared error; a general RKHS moment norm need not equal the projection norm used by kernel anchor regression (Shi & Xu, 2023).

To relate this identity to classical k-class estimation, put the instrument features in a matrix Q with $Q^\top Q/n = I_m$, and let $P_Q = QQ^\top/n$. Suppose also that their row norms are constant, $\|q_i\|^2 = m$. Then

$$\begin{aligned} \widehat{R}_\tau^2 &= \frac{1}{n} r^\top\{(1 - \tau)P_Q + \tau m I_n\}r \\ &= \frac{c_\tau}{n} r^\top\{I_n - k_\tau(I_n - P_Q)\}r, \qquad c_\tau = 1 - \tau + \tau m, \quad k_\tau = \frac{1 - \tau}{c_\tau}. \end{aligned} \tag{24}$$

With linear structural features and no ridge penalty, minimizing RMS therefore gives the corresponding k-class estimator whenever its normal matrix is invertible. For $0 \le \tau \le 1$, the path goes from projected least squares at $k_\tau = 1$ to ordinary least squares at $k_\tau = 0$. By Jakobsen & Peters (2022, Section 2.3), this is anchor regression with penalty $k_\tau/(1-k_\tau) = (1-\tau)/(\tau m)$, which in their linear model is minimax for prediction under anchor interventions of strength $1 + (1-\tau)/(\tau m)$; this strength decreases as τ grows. Nonconstant row norms replace mI_n by $\mathrm{diag}(\|q_1\|^2, \dots, \|q_n\|^2)$; a general moment metric also changes $QQ^\top/n$ from an orthogonal projector. These cases give a weighted operator analogue, rather than the same classical estimator. In particular, the HC0 metric used in the experiments does not by itself imply the whitening and constant-row-norm assumptions above.

Ridge requires care because we optimize a norm rather than its square. If the stacked system is $F_\tau(\theta) = c - K_\tau\theta$ and $R_\tau(\theta) = \|F_\tau(\theta)\| > 0$, stationarity of $R_\tau(\theta) + \lambda\|\theta\|^2/2$ gives

$$(K_\tau^\top K_\tau + \lambda R_\tau(\theta) I)\theta = K_\tau^\top c.$$

Thus its effective squared-loss ridge is $\lambda R_\tau(\theta)$, not λ. At zero residual the norm requires a subgradient instead.

Conversely, joint sensitivity can remove a null direction without displacing the truth. Let Z be constant, $q = 1$, let T and ε_0 be independent mean-zero Rademacher variables, and set $Y = \theta_0 T + \varepsilon_0$ with candidates $f_{\alpha,\beta}(T) = \alpha + \beta T$. The nominal CMR identifies only $\alpha = 0$, while full TAN gives

$$R_\tau(\alpha, \beta)^2 = \alpha^2 + \tau\big((\beta - \theta_0)^2 + 1\big).$$

For every $\tau > 0$, its unique minimizer is $(0, \theta_0)$ despite positive sensitivity at the truth. Here $c_q = 0$, and the preserving joint score $a = T$ supplies the missing slope moment. The example shows why displacement is not inevitable; validity of such additional identifying moments is not implied by the nominal IV restriction alone.

E From finite features to RKHS stability

To extend the finite-dimensional argument, assume separable RKHSs and

$$\mathbb{E}_P[k(Z,Z)\ell(T,T)] < \infty, \qquad \mathbb{E}_P[Y^2 k(Z,Z)] < \infty, \qquad \mathbb{E}_{P_w}[\ell(T,T)] < \infty.$$

For a closed separable score space $\mathcal{S} \subseteq L^2_0(P)$, define $\mathcal{B}_h : \mathcal{S} \to \mathcal{H}$ by $\mathcal{B}_h a = \mathbb{E}_P[a h(T) k(Z, \cdot)]$. These assumptions make $h \mapsto \mathcal{B}_h$ a bounded map into $\mathrm{HS}(\mathcal{S}, \mathcal{H})$, and make $J_w : \mathcal{F} \to L^2(P_{w,T})$ bounded. Indeed, $\|h(T)k(Z,\cdot)\|^2 \le \|h\|^2_{\mathcal{F}} \ell(T,T) k(Z,Z)$; Parseval's identity bounds the squared Hilbert–Schmidt norm by the expectation on the right. The same argument applies to the outcome response.

Write $\mathcal{K}_\tau h = (A_P h, \sqrt{\tau}\,\mathcal{B}_h)$, with norm in $\mathcal{H} \oplus \mathrm{HS}(\mathcal{S}, \mathcal{H})$. A uniform linear transfer from moment error to target error requires

$$\|J_w h\|_{L^2(P_{w,T})} \le c\|\mathcal{K}_\tau h\| \quad \text{for all } h \in \mathcal{F}. \tag{25}$$

When this bound fails, we can instead measure stability on an RKHS norm ball through

$$\omega_{2R}(t) = \sup\{\|J_w h\| : \|h\|_{\mathcal{F}} \le 2R,\ \|\mathcal{K}_\tau h\| \le t\}.$$

This modulus records how much target error can remain when the moments are small and structural complexity is bounded.

The stacked affine response is $\mathcal{F}_\tau(f) = (\mu_P(f), \sqrt{\tau}\,\mathcal{D}_f|_{\mathcal{S}})$, with norm in $\mathcal{H} \oplus \mathrm{HS}(\mathcal{S}, \mathcal{H})$. Its difference satisfies $\mathcal{F}_\tau(f+h) - \mathcal{F}_\tau(f) = -\mathcal{K}_\tau h$. For an infinite score space, this norm is a finite response-energy criterion under the stated assumptions; we do not posit an isotropic random L^2 score with covariance τI, which would have infinite total variance.

Proposition 6 (Identification and stability in an RKHS) *For $\tau > 0$,*

$$\ker \mathcal{K}_\tau = \ker A_P \cap \{h : \mathcal{B}_h a = 0 \text{ for every } a \in \mathcal{S}\}.$$

If the nominal and all retained score moments vanish at $f_0 \in \mathcal{F}$, the population RMS minimizers are $f_0 + \ker \mathcal{K}_\tau$. Moreover, equation 25 holds for some finite c if and only if there is a bounded linear map $L : \overline{\mathrm{ran}\,\mathcal{K}_\tau} \to L^2(P_{w,T})$ such that $J_w = L\mathcal{K}_\tau$. Under the stated integrability assumptions, J_w is compact. If $\ker \mathcal{K}_\tau \subseteq \ker J_w$, then $\omega_{2R}(t) \to 0$ as $t \downarrow 0$ for each finite R.

Proof. The null-space and minimizer statements follow by taking the squared norm of the stacked response, as in Proposition 4. If the stability bound holds, set $L(\mathcal{K}_\tau h) = J_w h$. The bound makes this well-defined, since $\mathcal{K}_\tau h = 0$ implies $J_w h = 0$, and gives $\|L\| \le c$ on the range. Continuity extends L to its closure. Conversely, such a factorization gives $\|J_w h\| \le \|L\|\|\mathcal{K}_\tau h\|$.

For an orthonormal RKHS basis (e_j), Parseval's identity gives $\sum_j \|J_w e_j\|^2 = \mathbb{E}_{P_w}[\ell(T,T)] < \infty$, so J_w is Hilbert–Schmidt and hence compact. If the modulus did not tend to zero, there would be a bounded sequence h_j with $\|\mathcal{K}_\tau h_j\| \to 0$ but $\|J_w h_j\|$ bounded away from zero. A subsequence converges weakly to some h; boundedness of $\mathcal{K}_\tau$ gives $\mathcal{K}_\tau h = 0$, and compactness gives $J_w h_j \to J_w h$. The null-space inclusion implies $J_w h = 0$, a contradiction.

In infinite dimensions, inclusion of the null spaces is necessary but insufficient for the uniform linear bound: on ℓ^2, take $Ke_j = j^{-2}e_j$ and $Je_j = j^{-1}e_j$. Both are compact and injective, yet no finite constant satisfies $\|Jh\| \le c\|Kh\|$. For finite features, writing $H = K_\tau^\top K_\tau$ and $\Sigma_w = \mathbb{E}_{P_w}[\phi\phi^\top]$, the sharp constant in equation 17 is

$$c_{w,\tau} = \lambda_{\max}(H^{\dagger/2}\Sigma_w H^{\dagger/2}),$$

provided $\ker H \subseteq \ker \Sigma_w$. This follows by substituting $h = H^{\dagger/2}v$ on the identified subspace into the quadratic-form ratio.

For a bound without uniform stability, assume the candidate class contains $f_\star$ and $\widehat{f}$, both of RKHS norm at most R. Suppose $\widehat{f}$ minimizes $\widehat{R}_\tau(f) + \lambda\|f\|_{\mathcal{F}}^2/2$ to accuracy ζ, and $|\widehat{R}_\tau(f) - R_\tau(f)| \le e_\tau(f)$ at these two functions. The proof of Proposition 5 gives

$$\|\mathcal{K}_\tau(\widehat{f} - f_\star)\| \le t_\star := 2R_\tau(f_\star) + e_\tau(f_\star) + e_\tau(\widehat{f}) + \lambda\|f_\star\|_{\mathcal{F}}^2/2 + \zeta.$$

The definition of ω_{2R} and the triangle inequality then imply

$$\|\widehat{f} - f_0\|_{L^2(P_{w,T})} \le \omega_{2R}(t_\star) + \|f_\star - f_0\|_{L^2(P_{w,T})}. \tag{26}$$

The last term accounts for structural approximation. Taking $f_\star = f_0$ recovers the main-text statement. The modulus inequality itself also holds without compactness or target identification, but then need not imply vanishing error. If a source or smoothness condition gives $\|J_w h\| \le C\|\mathcal{K}_\tau h\|^\nu \|h\|_{\mathcal{F}}^{1-\nu}$ for $0 < \nu \le 1$, it instead supplies the explicit bound $\omega_{2R}(t) \le C(2R)^{1-\nu}t^\nu$.

F A SAMPLING BOUND FOR FIXED FINITE FEATURES

The deterministic bound becomes a sampling statement once the stacked moment operator concentrates. To make this step explicit, condition on feature maps, score functions, and a moment transform chosen independently of n i.i.d. estimation observations. Fix $\tau \ge 0$ and set

$$M(V) = \begin{pmatrix} q(Z)[-\phi(T)^\top \;\; Y] \\ \sqrt{\tau}s_1(V)q(Z)[-\phi(T)^\top \;\; Y] \\ \vdots \\ \sqrt{\tau}s_d(V)q(Z)[-\phi(T)^\top \;\; Y] \end{pmatrix}, \qquad J_\tau = \mathbb{E}_P[M(V)].$$

Then $F_\tau(\theta) = J_\tau(\theta^\top, 1)^\top$ and its empirical counterpart uses $\widehat{J}_\tau = n^{-1}\sum_i M(V_i)$. If $\|M(V)\|_F \le B$ almost surely, then, with probability at least $1 - \delta$,

$$\|\widehat{J}_\tau - J_\tau\|_{\mathrm{op}} \le \epsilon_{n,\delta} := \frac{B}{\sqrt{n}}\big(1 + \sqrt{2\log(1/\delta)}\big). \tag{27}$$

To prove this, independence gives $\mathbb{E}\|\widehat{J}_\tau - J_\tau\|_F^2 \le B^2/n$. Changing one observation changes this Frobenius norm by at most $2B/n$; the bounded-differences inequality (McDiarmid, 1989) therefore adds at most $B\sqrt{2\log(1/\delta)/n}$ to its expectation with the stated probability. Finally, the operator norm is bounded by the Frobenius norm.

On this event, the reverse triangle inequality gives, simultaneously for all θ, $e_\tau(\theta) = \epsilon_{n,\delta}\sqrt{1 + \|\theta\|^2}$. For example, on a coefficient ball of radius R, the sum of the two estimation-error terms in Proposition 5 is at most $2\epsilon_{n,\delta}\sqrt{1 + R^2}$. The bound displays its dimension and tail dependence through

$$\|M(V)\|_F^2 = \|q(Z)\|^2(\|\phi(T)\|^2 + Y^2)(1 + \tau\|s(V)\|^2).$$

It covers fixed or independently fitted finite score dictionaries with this boundedness condition. It does not assert the same rate for unbounded outcomes, an empirical full-TAN score basis, or same-sample feature and HC0 metric learning; those require corresponding tail and dependence bounds.

G RESPONSE ENERGY AT THE STRUCTURAL TRUTH

The following calculation compares expected response energies at one candidate, not the strength of regularization away from that candidate. It uses fixed population-centered scores; empirical orthonormalization, learned features, and a same-sample residual-based metric require separate analysis.

Proposition 7 (Expected KS and TAN response energies) *Let $(T_i, Y_i, Z_i)_{i=1}^n$ be i.i.d., let $\varepsilon_i = Y_i - \phi(T_i)^\top \theta_0$ satisfy the CMR, and fix instrument features q and centered instrument scores $s \in \mathbb{R}^d$ with $\mathbb{E}_P[ss^\top] = I_d$. Assume*

$$a_0 := \mathbb{E}_P[\|q\|_2^2 \varepsilon_0^2] < \infty, \qquad b_0 := \mathbb{E}_P[\|s\|_2^2 \|q\|_2^2 \varepsilon_0^2] < \infty.$$

Define $\widehat{D}_{\mathrm{KS}} = n^{-1} \sum_i q_i \varepsilon_i s_i^\top$, and let $\widehat{D}_{\mathrm{TAN}}$ be the full empirical TAN response. Then

$$\mathbb{E}\|\widehat{D}_{\mathrm{KS}}\|_F^2 = b_0/n, \qquad \mathbb{E}\|\widehat{D}_{\mathrm{TAN}}\|_F^2 = (1 - 1/n) a_0.$$

If $b_0 > 0$, the ratio of the TAN and KS expectations is $(n-1)a_0/b_0$.

Proof. The CMR gives $\mathbb{E}[q\varepsilon_0 s^\top] = 0$ and $\mathbb{E}[q\varepsilon_0] = 0$. Independence makes the cross terms in the squared norm of the empirical KS average vanish, leaving b_0/n. By equation 22, full TAN energy is $n^{-1} \sum_i \|q_i\|_2^2 \varepsilon_i^2 - \|n^{-1} \sum_i q_i \varepsilon_i\|_2^2$. The two expectations are a_0 and a_0/n, proving the second identity.

Scaling proportional to n/d additionally requires b_0 to be comparable to da_0; orthonormality alone does not control the association of score leverage with $\|q(Z)\|_2^2 \mathbb{E}[\varepsilon_0^2 \mid Z]$. The proposition concerns a ratio of expectations, not the expectation or concentration of the observed ratio. At a candidate other than the truth, write $r_\theta = Y - \phi(T)^\top \theta$. Whenever the corresponding second moments are finite, the fixed-score identity instead reads

$$\mathbb{E}\|\widehat{D}_{\mathrm{KS}}(\theta)\|_F^2 = \|D_{\mathrm{KS}}(\theta)\|_F^2 + \frac{1}{n} \mathbb{E}\|q r_\theta s^\top - D_{\mathrm{KS}}(\theta)\|_F^2.$$

In particular, $D_{\mathrm{KS}}(\theta_0 + h) = -[A_1 h, \ldots, A_d h]$ can supply substantial population curvature even though its value vanishes at the truth. A pilot ratio $\|\widehat{D}_{\mathrm{TAN}}(\theta_{\mathrm{pilot}})\|_F^2 / \|\widehat{D}_{\mathrm{KS}}(\theta_{\mathrm{pilot}})\|_F^2$, when defined, matches squared response penalties at that pilot; it does not match their directions, curvature, distributional budgets, or confidence levels.

H MOMENT METRIC AND EXPERIMENTAL SELECTION

The score space and criterion specify the estimator once the features, moment metric, and penalty are fixed. Here we give the choices used for the reported comparisons, since changing the metric or the selection rule also changes the fitted estimator.

Moment metric. We standardize the training outcome and use the same Nyström representation (Williams & Seeger, 2001) of Gaussian kernels to form nominal and score moments. Let q_i^{raw} denote the instrument features before the moment transform. A sieve 2SLS pilot (see Baselines below) with penalty 10^{-3} gives residuals $\widetilde{r}_i$ on the training sample. We then form

$$C = \frac{1}{n}\sum_i \widetilde{r}_i^2 q_i^{\mathrm{raw}}(q_i^{\mathrm{raw}})^\top + 10^{-2} I, \qquad LL^\top = C, \qquad q_i = L^{-1} q_i^{\mathrm{raw}}.$$

This heteroskedasticity-consistent (HC0) transform (White, 1980) is held fixed during each criterion's optimization and applied to both its nominal and response moments. It uses the same sample as the fit, so the independent-feature assumptions of Appendix F are not a statistical guarantee for this pipeline.

Penalty selection. For our criteria, the grid consists of eleven logarithmically spaced penalties from 10^{-6} to 10^{-1}. The EB rule selects a penalty using TAN–RMS and shares it across the response criteria at that radius; in the benchmark and extended-radius runs, it is recomputed at each radius. To specify the rule, write the compact TAN–RMS system as $B_\tau = [-M_\tau, c_\tau]$, so that $\|c_\tau - M_\tau\theta\|_2 = \widehat{R}_\tau(\theta)$. The implementation stacks a thin QR root of the nominal affine moments and $\sqrt{\tau}$ times an eigendecomposition root of the full TAN response Gram matrix, retaining all columns and root rows, including zero eigenvalues. Let r be the number of rows of this stacked root, let $M_\tau = U\,\mathrm{diag}(s_j)V^\top$ be a thin SVD, and set $\alpha_\lambda = \max\{\lambda\|c_\tau - M_\tau\widehat{\theta}_\lambda\|_2, \epsilon\}$, where ϵ is machine precision. The selection score is

$$Q_\lambda = \|c_\tau\|_2^2 - \sum_j \frac{s_j^2}{s_j^2 + \alpha_\lambda}(U_j^\top c_\tau)^2,$$

$$\mathrm{EB}(\lambda) = r\log\big(\max\{Q_\lambda/r, \epsilon\}\big) + \sum_j \log(1 + s_j^2/\alpha_\lambda).$$

This is the evidence of Bayesian ridge regression with the noise variance profiled out (MacKay, 1992; Bishop, 2006), evaluated at the effective ridge α_λ of the unsquared criterion (Appendix D). We minimize this empirical-Bayes-inspired score on the specified compact representation; we do not assume a generative likelihood for these dependent moment rows. The row count matters: appending zero rows preserves the estimation objective but can change the selection score. Competitors use five-fold validation PRE, and the real-data table uses this rule for all methods. The matched-CV comparison in Appendix J helps distinguish performance of the complete pipeline from effects attributable to the response criterion alone.

Joint selection and evaluation. For shifted training, draw a pool of $N = 10n$ observations from the target distribution and form a standardized linear index s_i of Z, Y, or (X, Y). Choose t so the mean-one weights $w_i = \exp(ts_i)/(N^{-1}\sum_j \exp(ts_j))$ satisfy $N^{-1}\sum_i w_i^2 - 1 = \chi^2$ at the requested strength. Draw independent uniforms U_i and retain the n largest keys U_i^{1/w_i}, a weighted sample without replacement (Efraimidis & Spirakis, 2006). The reported strength is the divergence of these pool weights; the marginal inclusion probabilities of the selected sample are not exactly proportional to w_i. For a population training reference P_{train}, a full TAN ball would contain target P if $P \ll P_{\mathrm{train}}$ and $\rho^2 \geq \chi^2(P\|P_{\mathrm{train}}) < \infty$. This reverse divergence differs from the forward divergence describing a tilt away from P; the recorded pool-weight strength establishes neither population quantity. Structural MSE is evaluated on 2 000 independent target observations. For test-time tilts, training stays unshifted and the test moments are evaluated directly under normalized exponential weights, with the largest MMR reported over the tested directions and signs.

Real-world data. Airline uses U.S. domestic market data from the DB1B survey (BTS, 2025), with log fare as treatment, log passengers as outcome, and leave-market-out fares as instruments;

grouped splits hold out markets. Automobile uses the demand data of Berry et al. (1995) with the excluded demand instruments distributed with PyBLP (Conlon & Gortmaker, 2020); grouped splits hold out markets. JTPA uses the randomized training offer as instrument for participation, with earnings as outcome (Bloom et al., 1997); grouped splits hold out one sex. 401(k) uses plan eligibility as instrument for participation, with net financial assets as outcome (Abadie, 2003); grouped splits hold out five-year age bands.

I BENCHMARK DESIGNS

Our main design instantiates equation 16 with four treatments and four instruments. The instruments are independent and uniform, the first stage is linear with a first-stage R^2 set by rescaling the instrument signal, a shared Gaussian confounder loads equally on every treatment and on the outcome, and h_0 is a coordinatewise absolute-value function. The main-design variants change one of these components at a time: the treatment dimension, the structural function (a random tanh network or random Fourier features), heteroskedastic confounding, random confounder loadings, sparse mixing of eight instruments into four treatments, a nonlinear first stage, correlated instruments, the confounding share of treatment variation, or observation of (X, Z) through fixed random nonlinear embeddings in 32 to 128 dimensions. The invalid-instrument variants of the main design add a direct effect of the first instrument on Y, or a Gaussian-copula dependence between that instrument and the confounder. With image instruments, the observed instruments are the leading principal components of MNIST, Fashion-MNIST, or CIFAR-10 images (LeCun et al., 1998; Xiao et al., 2017; Krizhevsky, 2009), while the latent instrument driving the treatment is a fixed projection of all components; the pool rank and the first-stage strength are varied separately.

Four further designs follow prior work. The demand design of Kremer & Schölkopf (2024), after Hartford et al. (2017), is the special case of the main design with identity mixing and equal treatment and instrument dimensions; we use absolute-value and random-network structural functions in 1 to 10 dimensions. The linear design follows Qu & Kwon (2024, Section 5.1): $X = Z + U\mathbf{1}$ and $Y = \beta^\top X + \eta Z_1 + U$ with $\beta = \mathbf{1}/\sqrt{d}$, $d = 10$, and $Z = \beta_{UZ} U e_1 + \epsilon$; η violates exclusion and β_{UZ} exogeneity, and Fig. 4 uses it for both violation panels. The contamination design of Kremer & Schölkopf (2024) has a scalar instrument, treatment, and quadratic structural function, and replaces a fraction of training treatments by uniform draws over their range; Fig. 4 uses it for the corruption panel. The single-index attack design, after Kremer & Schölkopf (2024), drives a five-dimensional treatment with a scalar instrument and uses $h_0(t) = 1.5\cos(a^\top t) + 0.1 a^\top t$; we calibrate its identification strength and signal share like our other designs, since the published constants leave the index essentially unidentified. The attack panel of Fig. 4 uses the main design.

J ADDITIONAL RESULTS

Unless stated otherwise, our estimators use $\kappa = 3$ and the EB selection score from Appendix H, competitors use five-fold cross-validation (CV) of the projected residual, and benchmark results are reported as structural MSE divided by the best CV-tuned competitor on the same task, averaged over the tasks of an experiment. A value below one indicates a lower mean task-normalized error, not a win on every task.

Selection rule. Figure 5 gives our criteria the competitors' CV rule on the same eleven-point λ grid. It also reports TAN–Spectral and KS–Frobenius, which are omitted from the crowded main displays. On the benchmarks, EB is better for every TAN criterion in almost every experiment; on real data, where the evaluation criterion is itself the projected residual, CV is comparable or better.

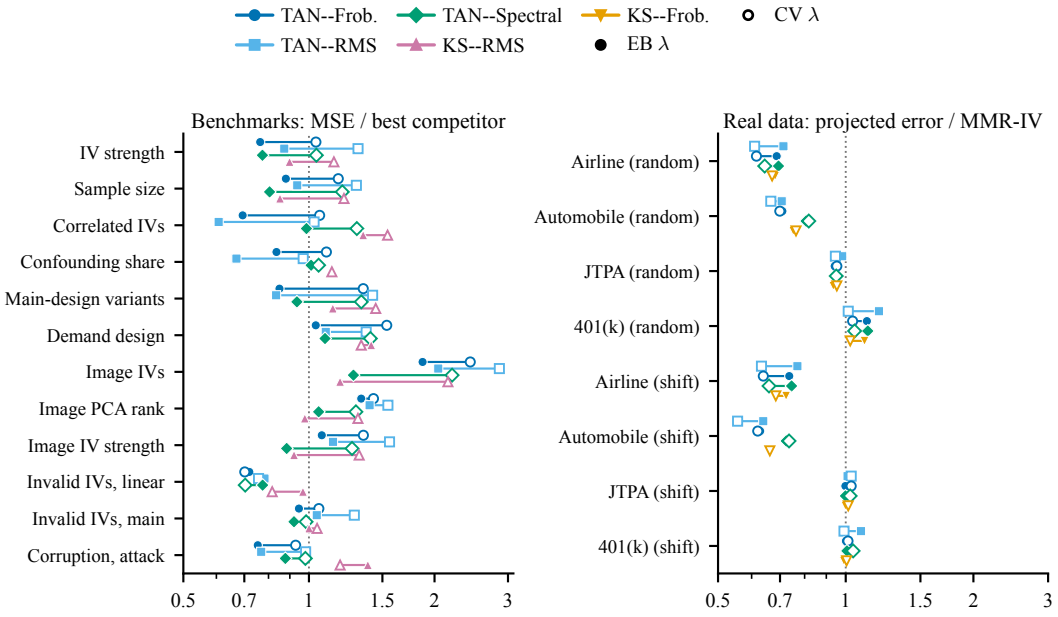


Figure 5: EB (filled) versus CV (open) penalty selection at $\kappa = 3$. Left: mean task-normalized structural MSE relative to the best competitor. Right: held-out projected residual error relative to MMR-IV on real data.

Radius paths, the regression endpoint, and KS units. Figure 6 extends the radius path to $\kappa = 10^4$. For unregularized TAN–RMS, increasing $\tau = \kappa/n$ from zero to one moves the squared objective from nominal moment loss to weighted residual regression (Proposition 2 and Appendix D). Beyond $\tau = 1$, TAN–RMS increasingly emphasizes moment-contribution variance, whose minimizer need not be a regression fit; neither this limit nor the RMS identity gives a regression limit for the upper envelopes. The KRR curve is therefore a descriptive comparison throughout the plotted path. Reweighting by functions of Z preserves $\mathbb{E}[\varepsilon_0 \mid Z] = 0$, so the population KS response vanishes at the truth, while its empirical response includes sampling error. Equal κ need not give comparable response penalties in KS and TAN; Appendix G gives their expected energies at the truth. Calibrated KS rescales its weight by $c = \|D_{\mathrm{TAN}}(\theta_{\mathrm{pilot}})\|_F^2 / \|D_{\mathrm{KS}}(\theta_{\mathrm{pilot}})\|_F^2$ (median c between 12 and 117 across experiments), matching response energy at one pilot rather than a common distributional budget. This moves its best radius into the range of TAN's. With the EB penalty, $\kappa = 0$ inherits a poorly regularized nominal fit.

Real-data metrics. Both held-out criteria carry the units of the outcome (squared for the projected error), which differ by orders of magnitude across datasets. Figure 7 reports scale-free per-seed ratios to MMR-IV, pairing methods on the same splits; for MMR, every method is also tuned on validation MMR. MMR separates the methods much less than the projected error does.

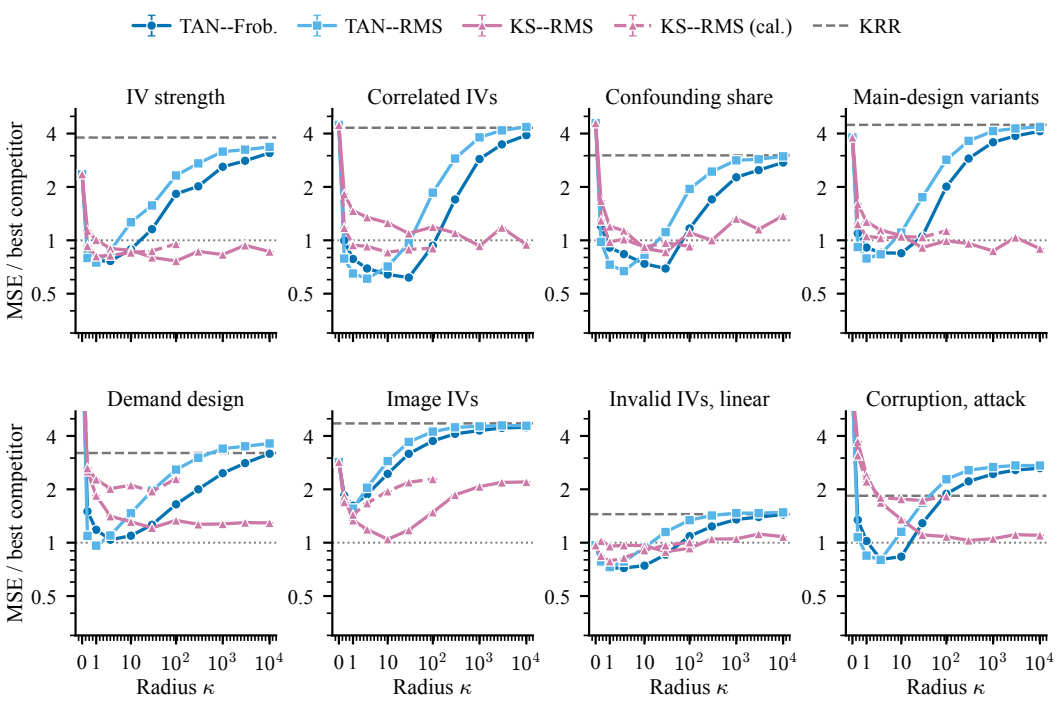


Figure 6: Structural MSE relative to the best competitor along the radius path (EB penalty) for TAN, KS, and calibrated KS (shown up to $\kappa = 100$). The dashed line is KRR, a regression of Y on X without instruments.

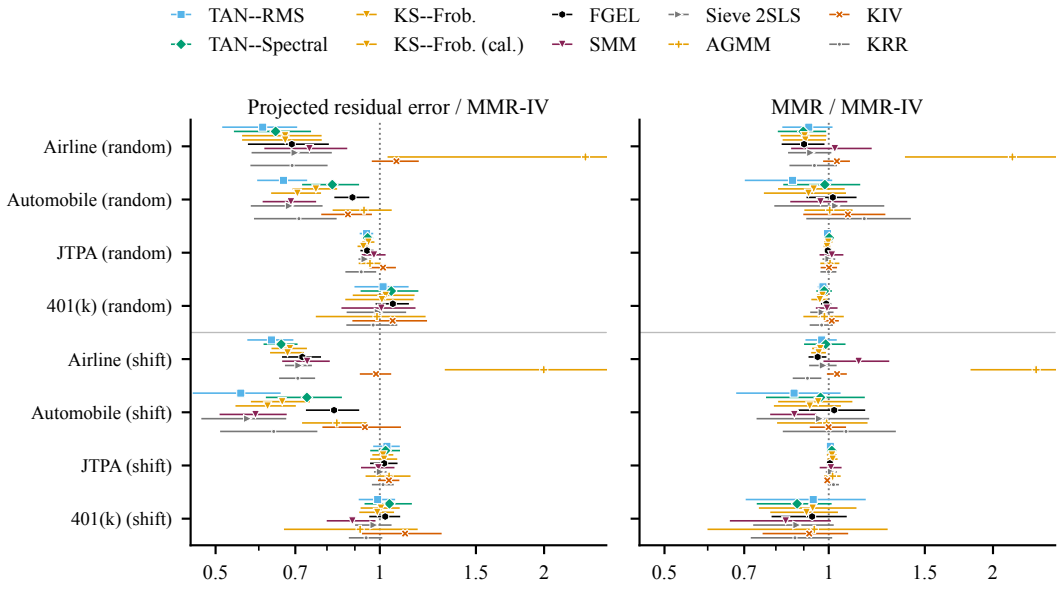


Figure 7: Held-out projected residual error (left) and MMR (right) relative to MMR-IV: mean and 95% interval of per-seed ratios over ten seeds, CV selection for all methods.

Joint training–target shifts. Figure 8 extends Fig. 3 to further designs. TAN retains the lowest absolute error in most cells, but its error grows faster with the shift than that of Sieve 2SLS or SMM. The structural-recovery advantage at small shifts therefore does not imply better relative stability as the shift grows.

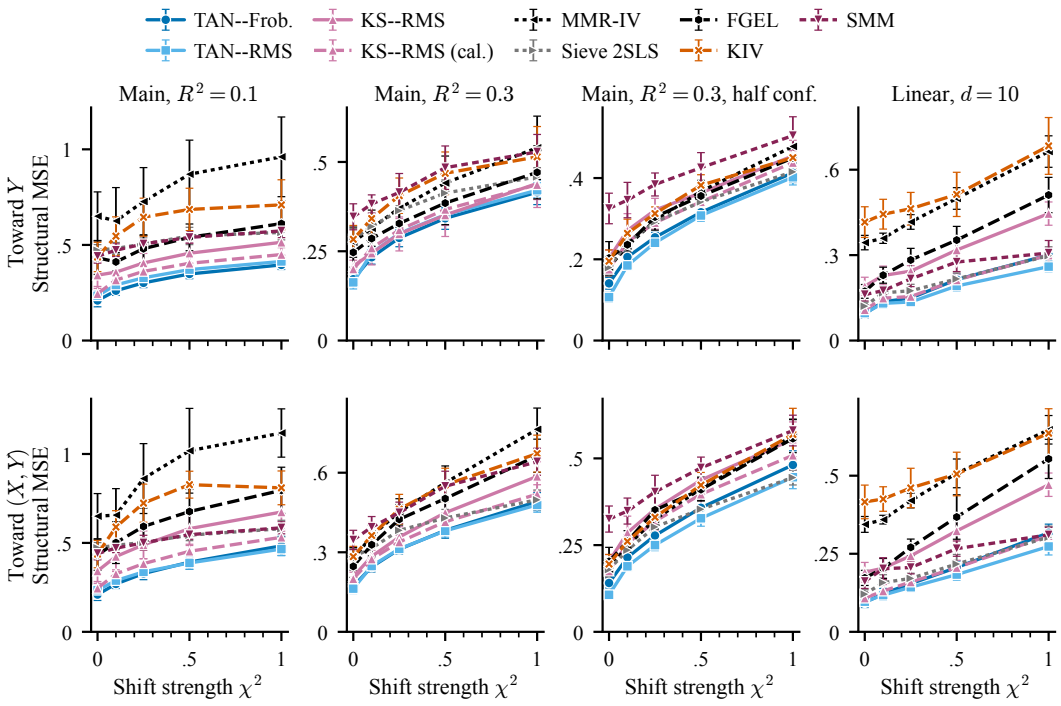


Figure 8: Structural MSE on the target population after weighted training-sample selection toward Y (top) or (X, Y) (bottom), against nominal pool-weight strength χ^2, for four designs.

Test-time tilts. Figure 9 keeps training unshifted and reweights the test sample, reporting the worst held-out MMR over directions and signs relative to Sieve 2SLS. Under joint tilts, even the structural function h_0 violates the moment condition, and KRR and large-radius TAN have the smallest violations; a small tilted MMR is therefore not a structural virtue. Under Z-only tilts, the condition holds at h_0, and TAN and calibrated KS at $\kappa = 3$ are among the methods closest to it.

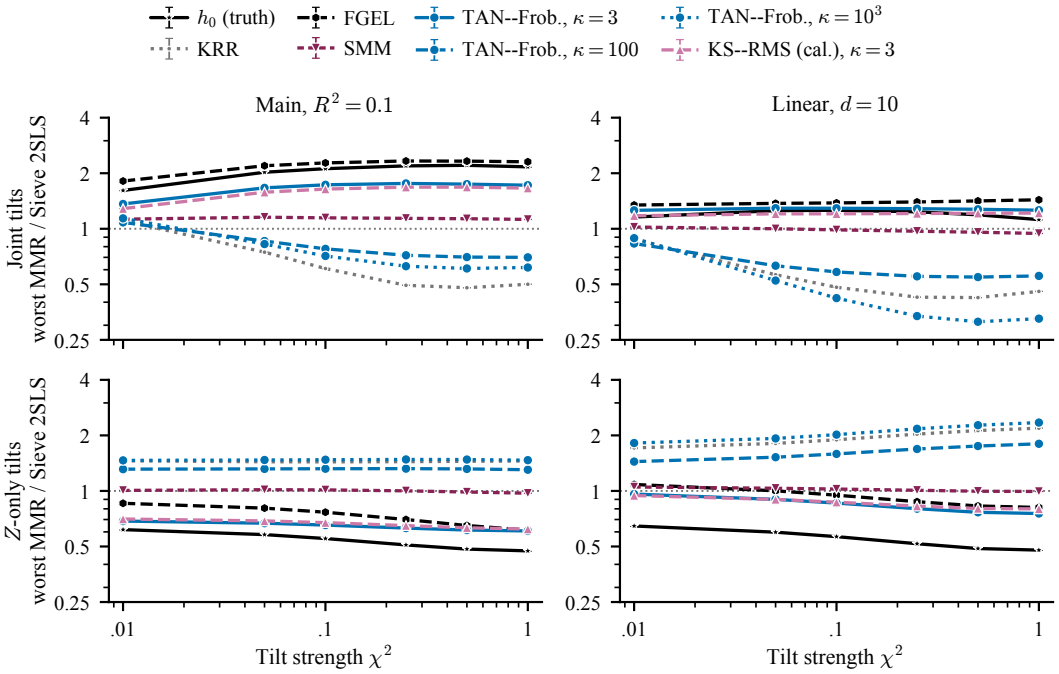


Figure 9: Worst held-out MMR under test-time tilts of strength χ^2, relative to Sieve 2SLS, for joint (X, Y, Z) directions (top) and instrument-only directions (bottom). Training is unshifted.