

POWER ENHANCEMENT VIA CROSS-FIT VARIANCE ESTIMATION: APPLICATIONS TO SPECIFICATION, OVERIDENTIFICATION, AND MANY-RESTRICTION TESTING

Keita Sunada^{a,*}, Yukitoshi Matsushita^b, and Taisuke Otsu^c

^a*Aarhus Center for Econometrics, Aarhus University, Universitetsbyen 51, 8000 Aarhus C, Denmark*

^b*Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan*

^c*Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom*

*Corresponding author. E-mail address: ksunada@econ.au.dk

ABSTRACT. Quadratic-form test statistics are widely used in econometrics, and their performance depends on accurate variance estimation. Conventional plug-in estimators are consistent under the null hypothesis, but under alternatives the drift in the residuals inflates them and the test loses power. We develop a general framework for variance estimation in such statistics, replacing one of the two squared-residual factors by an auxiliary linear combination of the residuals (“cross-fitting”) chosen to annihilate the drift. We characterize the conditional bias of each estimator exactly. The drift enters the plug-in estimator squared, multiplied by quantities bounded away from zero, so its bias is positive whenever the drift is non-degenerate. It reaches the cross-fit estimator only through the part that survives the cross-fitting, and then only through off-diagonal entries of a residual-maker matrix. From this calculation we obtain conditions under which the cross-fit estimator remains consistent under alternatives while the plug-in estimator does not. At a common critical value the cross-fit test therefore rejects whenever the plug-in test does with probability approaching one, and against distant alternatives the plug-in statistic converges to a finite limit, small when few observations carry the departure: its power can tend to zero where the cross-fit test’s tends to one. We verify the conditions under primitive assumptions in nonparametric specification testing, overidentification testing, and testing many linear restrictions, and illustrate the procedure on the Oregon Health Insurance Experiment.

JEL Classification: C12, C14, C26

Keywords: variance estimation, cross-fitting, power, quadratic-form statistics, specification testing, overidentification, many restrictions.

1. INTRODUCTION

Quadratic-form test statistics play a central role in econometrics and arise in a wide range of applications, including nonparametric specification testing (e.g. Hong and White, 1995;

Donald *et al.*, 2003; Y. Sun and Li, 2006), overidentification testing in instrumental variables models (e.g. Anatolyev and Gospodinov, 2011; Lee and Okui, 2012; Chao *et al.*, 2014), and tests of many linear restrictions (e.g. Cattaneo *et al.*, 2018; Anatolyev and S¸¸lvsten, 2023). In these procedures, statistical significance is determined by studentized test statistics, making accurate variance estimation essential for reliable inference. Variance estimators are commonly constructed by plugging residual-based estimates into the population variance formula, and are justified by their consistency under the null hypothesis. Each of these literatures has accordingly been built around valid inference under the null and the asymptotic null distribution.

What such an estimator does under alternatives is considerably less understood, and it can have important consequences for power. Under fixed or local alternatives, the drift component of the residuals enters the variance estimator and can generate substantial variance inflation. As a result, the studentized test statistic may suffer a considerable loss of power even when the underlying quadratic-form statistic itself exhibits a strong signal. This issue becomes particularly pronounced in modern settings where the signal accumulates over many observations or restrictions.

We construct variance estimators for quadratic-form statistics that remain valid under the null while being substantially less sensitive to deterministic drift under alternatives. The construction replaces one of the two squared-residual factors by an auxiliary linear combination of the residuals (“cross-fitting”) chosen to annihilate the drift. Because the drift then survives in only one of the two factors, it can reach the variance estimate only through a second-order covariance, whereas it enters the plug-in estimator directly. This is the construction that Mikusheva and L. Sun (2022) introduced for the jackknife Anderson–Rubin statistic in a linear instrumental-variables model with many weak instruments; the framework below shows that neither the construction nor the power gain it delivers is specific to that setting.

We first compute the conditional bias of each estimator under the alternative exactly. The two formulas differ in where the systematic component of the residual can appear. The drift reaches the plug-in estimator through products of diagonal quantities, which are bounded away from zero. In contrast, it reaches the cross-fit estimator only through the part of the systematic component that survives the cross-fitting, and, when that part vanishes, only through off-diagonal entries of a residual-maker matrix. The plug-in bias is in addition nonnegative, so it inflates the estimated variance, while the cross-fit bias need not be. From this calculation we obtain conditions on the weights and on the residual decomposition under which the cross-fit estimator is consistent for the asymptotic variance under both hypotheses

and the plug-in estimator is inconsistent under the alternative. The condition behind the cross-fit consistency asks that the weights of the quadratic form not be concentrated on a few pairs; the condition behind the plug-in inconsistency asks that the systematic component be non-degenerate relative to the weights.

Referred to a common critical value, the two studentized statistics differ asymptotically by the factor $\sqrt{1 + B_n/V}$, where B_n is the plug-in bias and V the asymptotic variance, so the cross-fit test rejects whenever the plug-in test does with probability approaching one. Also, along local alternatives $B_n/V \rightarrow 0$ and the two tests have the same limiting power. However, against a distant alternative, the plug-in bias grows as fast as the square of the numerator, so the plug-in statistic converges to a finite limit. That limit is invariant to the scale of the systematic component, and it is bounded by the number of observations on which that component is nonzero. When that bound falls below the critical value, the plug-in statistic never reaches it however large the departure, and the plug-in test has power tending to zero where the cross-fit test has power tending to one.

We verify the conditions of the general theory under primitive assumptions in three important econometric applications: nonparametric specification testing, overidentification testing in instrumental variables models, and testing many linear restrictions in linear regression. In each, the asymptotic results follow as direct consequences of the general theory. Although these testing problems arise in different econometric settings, in each of them the drift enters the plug-in estimator squared; what cross-fitting removes differs across the three, as Remark 5 sets out. Monte Carlo experiments demonstrate that the proposed procedure keeps size close to nominal over most of the range, turning mildly liberal only in the smallest-sample designs, and delivers substantial power gains relative to conventional plug-in variance estimators. The gains can be substantial when the number of restrictions or instruments is large relative to the sample size, and are largest for many restrictions, where the plug-in procedure also grows more conservative under the null as the number of restrictions increases. We further illustrate the practical relevance of the proposed approach through an empirical analysis of treatment effect heterogeneity in the Oregon Health Insurance Experiment.

Our work is related to the literature on cross-fitting, sample splitting, and leave-out estimation. Cross-fitting and sample splitting have become standard tools for eliminating overfitting bias and obtaining valid inference with flexible first-stage estimators (Newey and Robins, 2018; Chernozhukov *et al.*, 2018). Their use for variance estimation is a currently active area: leave-out and bias-corrected variance estimators have been developed for high-dimensional linear regression and quadratic-form inference (Kline *et al.*, 2020; Jochmans,

2022), for testing many linear restrictions under heteroskedasticity (Anatolyev and S¸olvsten, 2023), and more recently for quadratic forms under clustered sampling (Koles¸ar *et al.*, 2026).

Closest to this paper is Mikusheva and L. Sun (2022), who establish for the jackknife Anderson–Rubin statistic that the cross-fit estimator of the scale parameter remains consistent under local alternatives, and document the power loss that the plug-in estimator suffers. The present paper generalizes that result in four directions: general weighting arrays and auxiliary vectors in place of a single statistic, an estimated nuisance parameter in the residual, alternatives beyond their local-drift regime, and an exact conditional bias with a nonnegative lower bound for the plug-in estimator in place of an approximation. See Remark 2 for the details.

The obstacle the leave-out literature confronts is that an unbiased estimator of $\sigma_i^2\sigma_j^2$ built from leave-out residuals is not invariant to the parameterization of the regression, and the literature has taken three routes around it. First, Anatolyev and S¸olvsten (2023, Remark 3) accept the non-invariance, and pay for it with a leave-three-out construction. Second, Anatolyev and Korobka (2026) restore invariance by splitting the sample, at the cost of a bound on how many restrictions may be tested — $q/n \leq 1/4$ for the pairwise products.¹ We take a third route: impose the restriction under test, which removes the part of the contamination that the restriction identifies, and characterize what remains.² Anatolyev and S¸olvsten (2023, Remark 4) point in this direction, and we investigate it in the setting where imposing the restriction leaves only a fixed-dimensional parameter to estimate. Under this narrower setup the resulting test delivers moderately higher power than theirs in our simulations and avoids their leave-three-out construction, which makes it far cheaper to compute.

The paper is organized as follows. Section 2 develops the general theory, Section 3 the three applications, and Sections 4 and 5 the simulation and empirical evidence; the Appendix contains all proofs.

¹Jochmans (2022, Section 3) explains why the first two routes diverge: an invariant cross-fit estimator needs two disjoint subsamples that each identify the parameter, which forces a bound on the number of restrictions.

²See also Boot (2023) for a different route to power under many restrictions: joint confidence regions centered at James–Stein averaging estimators, which gain power against sparse alternatives when the restricted estimator points at the right coefficients and lose it when it points at the wrong ones. Ours asks for no such choice, since we change only the variance estimator and leave the statistic alone, and Boot’s heteroskedasticity-robust variant uses the same leave-out variance estimator that the present paper replaces.

2. GENERAL RESULT

Let $\{\hat{u}_i\}_{i=1}^n$ denote estimated residuals or moment conditions. We consider statistics of the form

$$Q_n = \sum_{i=1}^n \sum_{j \neq i}^n a_{ij} \hat{u}_i \hat{u}_j, \quad (1)$$

where $\{a_{ij}\}$ is an array of weights that may depend on underlying variables. Such statistics arise in a variety of econometric applications, including specification testing, overidentification testing in instrumental variable models, and tests involving many restrictions. We suppose that $\hat{u}_i$ admits the decomposition

$$\hat{u}_i = \Delta_i + e_i + \rho_i, \quad (2)$$

where Δ_i denotes the leading drift component under the alternative hypothesis, e_i is a mean-zero stochastic error, and ρ_i is an asymptotically negligible remainder. Under the null hypothesis, $\Delta_i = 0$. Under the alternative hypothesis, the drift component may be non-negligible and can materially affect the behavior of variance estimators.

Let $\sigma_i^2 = \mathbb{E}[e_i^2]$ and

$$V = 2 \sum_{i=1}^n \sum_{j \neq i}^n a_{ij}^2 \sigma_i^2 \sigma_j^2 \quad (3)$$

denote the leading asymptotic variance component of Q_n . A conventional plug-in estimator of V is

$$\hat{V} = 2 \sum_{i=1}^n \sum_{j \neq i}^n a_{ij}^2 \hat{u}_i^2 \hat{u}_j^2. \quad (4)$$

Substituting (2) into (4) introduces drift-dependent terms under the alternative hypothesis. In particular, terms of the form $\Delta_i^2 \sigma_j^2$ and $\Delta_i^2 \Delta_j^2$ inflate the estimated variance and attenuate the resulting studentized statistic.

To reduce this contamination, we consider a cross-fit variance estimator. Let $\Delta = (\Delta_1, \dots, \Delta_n)'$, $e = (e_1, \dots, e_n)'$, $\rho = (\rho_1, \dots, \rho_n)'$, and ℓ_i be an auxiliary vector constructed so that $\ell_i' \Delta = 0$. We propose the following cross-fit variance estimator:

$$\hat{V}_{\text{cf}} = \sum_{i=1}^n \sum_{j \neq i}^n c_{ij} \{\hat{u}_i(\ell_i' \hat{u})\} \{\hat{u}_j(\ell_j' \hat{u})\}, \quad (5)$$

where the weights $\{c_{ij}\}$ are chosen to remove the bias. Let L be the matrix with rows ℓ_i' . If the e_i are independent with mean zero, then for $i \neq j$,

$$\mathbb{E}[e_i(\ell_i' e) e_j(\ell_j' e)] = (L_{ii} L_{jj} + L_{ij} L_{ji}) \sigma_i^2 \sigma_j^2,$$

so that taking

$$c_{ij} = \frac{2a_{ij}^2}{L_{ii}L_{jj} + L_{ij}L_{ji}}$$

makes $\mathbb{E}[\hat{V}_{\text{cf}}] = V$ exactly when $\Delta = \rho = 0$.

The two estimators differ in where the drift can appear. By squaring the residuals, $\hat{V}$ carries Δ_i^2 , which is nonnegative, accumulates over i , and cannot cancel. In contrast, pairing $\hat{u}_i$ with $\ell_i' \hat{u}$ removes that square, since $\ell_i' \Delta = 0$; products $\Delta_i \Delta_j$ survive for $i \neq j$, each carrying an off-diagonal weight built from L . In the applications, L is a residual-maker matrix $I - P$, so these weights are bounded by the maximal leverage $\max_i P_{ii}$. Lemma 2 makes the comparison exact.

In applications, we decompose ρ_i as

$$\rho_i = \omega_i + H_i' \hat{\delta}, \tag{6}$$

where ω_i is an approximation error that does not involve the estimation error, $H = (H_1, \dots, H_n)'$ is an $n \times p$ matrix, and $\hat{\delta}$ is an estimation error with $\|\hat{\delta}\| = O_p(r_n)$ for a sequence $r_n \rightarrow 0$. How large ω_i is depends on the quality of the approximation, and how large $H_i' \hat{\delta}$ is depends on the convergence rate r_n of the estimator.

Conditioning. The arrays $\{a_{ij}\}$ and $\{c_{ij}\}$ and the vectors $\{\ell_i\}$ may depend on underlying variables, typically covariates. Let $\mathcal{F}$ denote the σ -field generated by these variables together with Δ and ω . We assume that the errors are independent across i conditionally on $\mathcal{F}$ with $\mathbb{E}[e_i \mid \mathcal{F}] = 0$, and every expectation, order of magnitude and limit below is conditional on $\mathcal{F}$. In each application $\mathcal{F}$ is generated by the covariates and instruments, which makes the approximation error ω measurable. When the design is random, as it is in two of the three applications, conditions stated below as $o(\cdot)$ or $O(\cdot)$ are to be understood as holding in probability, and the conclusions then hold in probability as well. The matrix H need not be $\mathcal{F}$ -measurable. For example, in the overidentification application below, $H_i = x_i$ contains the first-stage error, which is correlated with e_i by construction. It enters only through the stochastic-order conditions on the estimation error.

Notation. We take the weights to be symmetric, $a_{ij} = a_{ji}$, as they are in all three applications. This is a normalization rather than a restriction because replacing a_{ij} by $(a_{ij} + a_{ji})/2$ leaves Q_n unchanged. Throughout, $\|\cdot\|$ is the Euclidean norm for vectors and the spectral norm for matrices.

The following assumption is maintained throughout this section. Let $\omega = (\omega_1, \dots, \omega_n)'$, $A_n = \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2$, and $C_n = \sum_{i=1}^n \sum_{j \neq i} |c_{ij}|$.

Assumption 1. The variables $\{e_i\}_{i=1}^n$ are, conditionally on $\mathcal{F}$, independent mean-zero random variables satisfying $c \leq \sigma_i^2 \leq C$ and $\max_{1 \leq i \leq n} \mathbb{E}[e_i^4] \leq C$ for some constants $0 < c < C < \infty$. Moreover, the leverage condition $L_{ii}L_{jj} + L_{ij}L_{ji} \geq c_L > 0$ holds for all $i \neq j$, so that $C_n \leq (2/c_L)A_n$, and there exists a deterministic sequence $b_n \geq 1$ such that

$$\begin{aligned} \max_{1 \leq i \leq n} |e_i| &= O_p(b_n), & \max_{1 \leq i \leq n} |\ell'_i e| &= O_p(b_n), \\ \max_{1 \leq i \leq n} |\omega_i| + \max_{1 \leq i \leq n} |\ell'_i \omega| &= o_p(b_n^{-3}), & \|\hat{\delta}\| &= O_p(r_n). \end{aligned}$$

The assumption requires ω to be small uniformly in i , but requires only a rate for $\hat{\delta}$. By (6) the estimation error enters into both factors of $\hat{u}_i(\ell'_i \hat{u})$: it enters $\hat{u}_i$ through $H'_i \hat{\delta}$ and $\ell'_i \hat{u}$ through $\ell'_i H \hat{\delta}$. Each factor is therefore of first degree in $\hat{\delta}$ and their product is of second degree, so $\hat{V}_{\text{cf}}$, which pairs two such products, is a polynomial of degree four whose coefficients are sums over $i \neq j$. The next lemma writes that polynomial explicitly.

Lemma 1. Let $\check{u}_i = \Delta_i + e_i + \omega_i$ be the residual stripped of estimation error, $\check{u} = (\check{u}_1, \dots, \check{u}_n)'$, and $h_i = \ell'_i H$, and define

$$g_{0i} = \check{u}_i(\ell'_i \check{u}), \quad g_{1i} = \check{u}_i h'_i + (\ell'_i \check{u}) H_i, \quad g_{2i} = \frac{1}{2}(H_i h_i + h'_i H'_i).$$

Then $\hat{u}_i(\ell'_i \hat{u}) = g_{0i} + g'_{1i} \hat{\delta} + \hat{\delta}' g_{2i} \hat{\delta}$ and

$$\hat{V}_{\text{cf}} = \sum_{i=1}^n \sum_{j \neq i} c_{ij} g_{0i} g_{0j} + \mathcal{G}'_1 \hat{\delta} + \hat{\delta}' \mathcal{G}_2 \hat{\delta} + \mathcal{R}_n, \quad (7)$$

where

$$\mathcal{G}_1 = \sum_{i=1}^n \sum_{j \neq i} c_{ij} (g_{0i} g_{1j} + g_{0j} g_{1i}), \quad \mathcal{G}_2 = \sum_{i=1}^n \sum_{j \neq i} c_{ij} (g_{1i} g'_{1j} + g_{0i} g_{2j} + g_{0j} g_{2i}),$$

and the remainder $\mathcal{R}_n$, which collects the terms of degree three and four in $\hat{\delta}$, satisfies $|\mathcal{R}_n| \leq \|\hat{\delta}\|^3 \mathcal{G}_3 + \|\hat{\delta}\|^4 \mathcal{G}_4$ for

$$\mathcal{G}_3 = \sup_{\|\delta\|=1} \left| \sum_{i=1}^n \sum_{j \neq i} c_{ij} \{ (g'_{1i} \delta)(\delta' g_{2j} \delta) + (g'_{1j} \delta)(\delta' g_{2i} \delta) \} \right|, \quad \mathcal{G}_4 = \sup_{\|\delta\|=1} \left| \sum_{i=1}^n \sum_{j \neq i} c_{ij} (\delta' g_{2i} \delta)(\delta' g_{2j} \delta) \right|.$$

If $\ell'_i H = 0$ for every i , then $g_{2i} = 0$, $g_{1i} = (\ell'_i \check{u}) H_i$, and $\mathcal{R}_n = 0$.

Under the law of large numbers (8) for the errors and the negligibility conditions (9) for the estimation error, the cross-fit estimator is consistent under the null. Both are verified for each application considered in Section 3.

Theorem 1. Suppose Assumption 1 holds and, under H_0 , (2) holds with $\Delta_i = 0$. Assume further that

$$\sum_{i=1}^n \sum_{j \neq i} \{c_{ij} e_i(\ell'_i e) e_j(\ell'_j e) - 2a_{ij}^2 \sigma_i^2 \sigma_j^2\} = o_p(V). \quad (8)$$

In addition, the estimation error is negligible for the variance:

$$\|\hat{\delta}\| \|\mathcal{G}_1\| = o_p(V), \quad \|\hat{\delta}\|^2 \|\mathcal{G}_2\| = o_p(V), \quad \|\hat{\delta}\|^3 \mathcal{G}_3 = o_p(V), \quad \|\hat{\delta}\|^4 \mathcal{G}_4 = o_p(V), \quad (9)$$

where $\mathcal{G}_1$ and $\mathcal{G}_2$ are defined in Lemma 1. Then

$$\hat{V}_{\text{cf}} = V + o_p(V).$$

Under the alternative the residual has a systematic component. Denote $\pi_i = \Delta_i + \omega_i$ for that systematic component and $\lambda_i = \ell'_i \omega$ for the part of it that survives the auxiliary factor ℓ'_i . Then the two factors of $\hat{u}_i(\ell'_i \hat{u})$ can be written as

$$\hat{u}_i = \underbrace{\Delta_i + \omega_i}_{\pi_i} + e_i + H_i \hat{\delta}, \quad \ell'_i \hat{u} = \underbrace{\ell'_i \Delta}_0 + \underbrace{\ell'_i \omega}_{\lambda_i} + \ell'_i e + \ell'_i H \hat{\delta},$$

the drift dropping out of the second because $\ell'_i \Delta = 0$. Let $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ and $\mu_{3i} = \mathbb{E}[e_i^3]$. In this notation the residual of Lemma 1 is $\tilde{u}_i = \pi_i + e_i$, and we write

$$\check{V} = 2 \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \tilde{u}_i^2 \tilde{u}_j^2, \quad \check{V}_{\text{cf}} = \sum_{i=1}^n \sum_{j \neq i} c_{ij} \{\tilde{u}_i(\ell'_i \tilde{u})\} \{\tilde{u}_j(\ell'_j \tilde{u})\},$$

which are (4) and (5) with $\hat{u}$ replaced by $\tilde{u}$ (i.e., $\hat{\delta} = 0$) throughout.

Lemma 2. Suppose (2) holds under H_1 with $\ell'_i \Delta = 0$, and let $\{e_i\}$ be independent with mean zero and bounded third moments. Then

$$B_n := \mathbb{E}[\check{V}] - V = 2 \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \{\pi_i^2 \sigma_j^2 + \sigma_i^2 \pi_j^2 + \pi_i^2 \pi_j^2\}, \quad (10)$$

and

$$\begin{aligned} B_n^{\text{cf}} := \mathbb{E}[\check{V}_{\text{cf}}] - V = & \sum_{i=1}^n \sum_{j \neq i} c_{ij} \left\{ \pi_i \lambda_i \pi_j \lambda_j + \pi_i \lambda_i L_{jj} \sigma_j^2 + L_{ii} \sigma_i^2 \pi_j \lambda_j \right. \\ & + \pi_i \pi_j (L \Sigma L')_{ij} + \pi_i \lambda_j L_{ij} \sigma_j^2 + \lambda_i \pi_j L_{ji} \sigma_i^2 \\ & \left. + \pi_i L_{ij} L_{jj} \mu_{3j} + \pi_j L_{ji} L_{ii} \mu_{3i} \right\}. \end{aligned} \quad (11)$$

The plug-in squares $\hat{u}_i$, so the systematic component multiplies itself and enters the bias as π_i^2 . The cross-fit multiplies $\hat{u}_i$ by $\ell'_i \hat{u}$, from which $\ell'_i \Delta = 0$ has already removed the drift; the systematic component of that factor is λ_i , and the bias carries $\pi_i \lambda_i$.

Every term of (10) is nonnegative, so with $\sigma_i^2 \geq c$ the plug-in bias is bounded below by $4c \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \pi_i^2$ and is strictly positive whenever $\sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \pi_i^2 > 0$. The cross-fit bias need not be positive, since its terms carry $\pi_i \lambda_i$ and μ_{3i} , whose signs are unrestricted, so no analogous lower bound is available.

The first two terms of (10) are quadratic in π and the last is quartic. The quadratic pair dominates for small π : with $\pi_i \equiv \pi$ and $\sigma_i \equiv \sigma$ they contribute $2\pi^2 \sigma^2$ against π^4 , so the two are comparable at $\pi^2 = 2\sigma^2$. Under local alternatives, along which $\pi_i \rightarrow 0$, the quartic term is the first to vanish and the quadratic one governs.

Remark 1. When $\ell'_i \pi = 0$ for every i so that $\lambda \equiv 0$, the bias reduces to $\sum_{i=1}^n \sum_{j \neq i} c_{ij} \{ \pi_i \pi_j (L \Sigma L')_{ij} + \pi_i L_{ij} L_{jj} \mu_{3j} + \pi_j L_{ji} L_{ii} \mu_{3i} \}$. Every surviving term carries an off-diagonal factor, $(L \Sigma L')_{ij}$ or L_{ij} , and the two third-moment terms, being linear rather than quadratic in π , dominate when the drift is small. The maintained assumption bounds $L_{ii} L_{jj} + L_{ij} L_{ji}$ away from zero but places no upper bound on the off-diagonal entries, so this reduction by itself does not make the bias negligible. Below, we impose a condition on those entries explicitly. ■

The next theorem is the main theoretical result of the paper. Under the alternative the cross-fit variance estimator remains consistent for V while the plug-in estimator converges to $V + B_n$, and B_n is bounded away from zero relative to V under a non-degeneracy condition on the systematic component.

Theorem 2. *Suppose Assumption 1 holds and, under H_1 , (2) holds with $\ell'_i \Delta = 0$. Let $\check{A}_i = \check{u}_i(\ell'_i \check{u})$ and assume that*

$$\begin{aligned} \sum_{i=1}^n \sum_{j \neq i} c_{ij} \{ \check{A}_i \check{A}_j - \mathbb{E}[\check{A}_i \check{A}_j] \} &= o_p(V), \\ 2 \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \{ \check{u}_i^2 \check{u}_j^2 - \mathbb{E}[\check{u}_i^2] \mathbb{E}[\check{u}_j^2] \} &= o_p(V + B_n), \\ \hat{V}_{\text{cf}} - \check{V}_{\text{cf}} &= o_p(V), \quad \hat{V} - \check{V} = o_p(V + B_n), \quad B_n^{\text{cf}} = o(V). \end{aligned} \tag{12}$$

Then

$$\frac{\hat{V}_{\text{cf}}}{V} \xrightarrow{p} 1, \quad \frac{\hat{V}}{V + B_n} \xrightarrow{p} 1.$$

If in addition $\liminf_n A_n^{-1} \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \pi_i^2 > 0$, then $\liminf_n B_n/V > 0$, so the plug-in estimator is not consistent for V .

The first two conditions of (12) are laws of large numbers for $\check{V}_{\text{cf}}$ and $\check{V}$; the next two control the estimation error, and are the alternative-hypothesis counterparts of (9); the last asks that the cross-fit bias be negligible. By Lemma 1, $\hat{V}_{\text{cf}} - \check{V}_{\text{cf}} = \mathcal{G}'_1 \hat{\delta} + \hat{\delta}' \mathcal{G}_2 \hat{\delta} + \mathcal{R}_n$, so

that the difference is $o_p(V)$ whenever $\|\hat{\delta}\|\|\mathcal{G}_1\| = o_p(V)$, $\|\hat{\delta}\|^2\|\mathcal{G}_2\| = o_p(V)$, $\|\hat{\delta}\|^3\mathcal{G}_3 = o_p(V)$, and $\|\hat{\delta}\|^4\mathcal{G}_4 = o_p(V)$. Its plug-in counterpart $\hat{V} - \check{V} = o_p(V + B_n)$ is the same statement for $\hat{V}$. When $\ell'_i H = 0$ the cross-fit expansion has $\mathcal{R}_n = 0$, so only $\mathcal{G}_1$ and $\mathcal{G}_2$ need bounding. However, the plug-in expansion keeps its degree-three and degree-four terms, which must be bounded as well. Lemma 3 gives aggregate bounds that imply the first two conditions of (12) and $B_n^{\text{cf}} = o(V)$, which the three applications verify below.

Remark 2. Mikusheva and L. Sun (2022) propose the jackknife Anderson–Rubin statistic in a linear instrumental-variables model with many weak instruments; their Theorem 3 shows that the cross-fit estimator of the scale parameter remains consistent under local alternatives, and their Section 4.2 documents the resulting power loss of the plug-in estimator. Theorem 2 extends that result in four directions. First, it covers general weighting arrays $\{a_{ij}\}$ and auxiliary vectors $\{\ell_i\}$ rather than a single statistic. Second, it allows the residual to depend on an estimated nuisance parameter through $H'_i \hat{\delta}$, which their implied error does not. The terms $\mathcal{G}_1, \dots, \mathcal{G}_4$ of Lemma 1 control that dependence. Third, it asks only that the cross-fit bias be negligible, in place of the upper bound on the drift relative to A_n that their local-alternative condition imposes. Finally, it establishes the inconsistency of the plug-in estimator formally; their Section 4.2 argues the point informally. The comparison extends to the bias expressions. Their leading term for the plug-in estimator is quartic in the drift alone, where (10) adds a quadratic one that dominates it at moderate drift. For the cross-fit estimator they establish consistency but no bias formula, where (11) is exact and has a third-moment term that leads when the drift is small. ■

Write $T_n = Q_n/\sqrt{\hat{V}}$ and T_n^{cf} for the same quantity with $\hat{V}_{\text{cf}}$ in the denominator. The next result shows that, compared against a common critical value, the two-sided test based on T_n^{cf} has a rejection region containing that based on T_n , and its rejection probability is at least as large.

Corollary 1. *Under the conditions of Theorem 2,*

$$\frac{T_n^{\text{cf}}}{T_n} = \sqrt{\frac{\hat{V}}{\hat{V}_{\text{cf}}}} = \sqrt{1 + \frac{B_n}{V}} \{1 + o_p(1)\}.$$

If in addition $\liminf_n B_n/V \geq \gamma > 0$, then $P(\hat{V} \geq \hat{V}_{\text{cf}}) \rightarrow 1$, so that $|T_n^{\text{cf}}| \geq |T_n|$ with probability approaching one.

Corollary 1 compares the two procedures at a common critical value. Under size correction the comparison need not favor either procedure. The simulations favor cross-fitting either

way: at a common critical value in Section 4, and after calibration in Section 5.3 and Appendix E, where most of the gap remains and it closes only against the smoother alternative at the smallest samples.

Remark 3. By Corollary 1 the gain from cross-fitting is governed by B_n/V , and this ratio depends on how the systematic component is distributed across observations, not only on its size. Dividing (10) by (3) gives

$$\frac{B_n}{V} = \frac{2 \sum_{i=1}^n \pi_i^2 \sum_{j \neq i} a_{ij}^2 \sigma_j^2}{\sum_{i=1}^n \sigma_i^2 \sum_{j \neq i} a_{ij}^2 \sigma_j^2} + \frac{\sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \pi_i^2 \pi_j^2}{\sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \sigma_i^2 \sigma_j^2}.$$

In the first term π_i^2 and σ_i^2 carry the same weight $\sum_{j \neq i} a_{ij}^2 \sigma_j^2$: two systematic components with the same $\sum_i \pi_i^2$ produce different gains, with a larger gain when the large values of π_i^2 belong to observations whose weight is large. ■

Along local sequences, those with $\pi_i \rightarrow 0$ sufficiently fast that $B_n/V \rightarrow 0$, the two statistics agree to first order and the two tests have the same limiting power. The comparison sharpens as the alternative moves away from the null. Write

$$N_n = \sum_{i=1}^n \sum_{j \neq i} a_{ij} \pi_i \pi_j, \quad S_n = 2 \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \pi_i^2 \pi_j^2, \quad \tau_n = \frac{N_n}{S_n^{1/2}},$$

for the systematic part of the numerator of Q_n , the quartic part of the plug-in bias (10), and their ratio. Two properties of τ_n follow from the definitions. Both N_n and $S_n^{1/2}$ are of second degree in π , so τ_n does not change when π is replaced by $c\pi$ for any $c \neq 0$. And if at most $k \geq 2$ of the π_i are nonzero, only the $k(k-1)$ pairs inside their support contribute, so Cauchy–Schwarz gives $|\tau_n| \leq \{k(k-1)/2\}^{1/2}$. The next result shows that the plug-in statistic converges to τ_n .

Corollary 2. *Suppose the conditions of Theorem 2 hold under H_1 , and in addition that (i) $Q_n = N_n + o_p(S_n^{1/2})$ and (ii) the quartic part of the plug-in bias dominates, $V + B_n = S_n\{1 + o(1)\}$. Then*

$$T_n = \{\tau_n + o_p(1)\} \{1 + o_p(1)\}, \quad \frac{T_n^{\text{cf}}}{T_n} \xrightarrow{p} \infty.$$

Let the test reject when $|T_n|$ exceeds a fixed critical value $c > 0$. If $\limsup_n |\tau_n| < c$ then $P(|T_n| > c) \rightarrow 0$; if in addition $\liminf_n |\tau_n| > 0$ then $P(|T_n^{\text{cf}}| > c) \rightarrow 1$.

Conditions (i) and (ii) hold along sequences of alternatives that move away from the null in a fixed direction. Write $\pi = t\pi^0$ for a fixed vector π^0 and a scalar $t \rightarrow \infty$. Then S_n is of order t^4 , while V does not depend on t , the quadratic part of (10) is of order t^2 , and the

TABLE 1. The three applications in the notation of Section 2. In each case $\ell_i = m_i$ is the i th column of M and $c_{ij} = 2a_{ij}^2/(M_{ii}M_{jj} + M_{ij}^2)$. The upper block maps each application into the general framework; the lower block highlights the application-specific components. Wherever $\ell_i' H = 0$, $g_{2i} = 0$ and $\mathcal{R}_n = 0$ in Lemma 1.

	Specification (Section 3.1)	Overidentification (Section 3.2)	Many restrictions (Section 3.3)
a_{ij}	$K^{-1/2}P_{ij}$	P_{ij}	Q_{ij}
M	$I - B(B'B)^{-1}B'$	$I - Z(Z'Z)^{-1}Z'$	$I - X(X'X)^{-1}X'$
Δ	$B\alpha_* - X\gamma_*$	$P\bar{u}$	$M_W Z(\beta_2 - \beta_{20})$
H_i	x_i	x_i	w_i
ω_i	$\theta_0(x_i) - \alpha_*' b(x_i)$	$m_i' \bar{u}$	0
$\lambda_i = m_i' \omega$	$o_p(b_n^{-3})$	ω_i	0
$\ell_i' H$	0	$m_i' v$	0

conditional variance of $Q_n - N_n$, which is $V + 4 \sum_{j=1}^n \sigma_j^2 (\sum_{i \neq j} a_{ij} \pi_i)^2$ up to the estimation error, is of order t^2 as well.

The same saturation arises for the jackknife Anderson–Rubin statistic in Mikusheva and L. Sun (2022, Section 4.2), where the bias of the analogous plug-in variance is quartic in the departure.

3. APPLICATIONS

This section applies the general framework of Section 2 to three representative testing problems: specification testing, overidentification testing in instrumental variable models, and testing many linear restrictions in linear regression. Table 1 collects the correspondence.

3.1. Specification testing. We first consider the series-based specification test of Y. Sun and Li (2006), which is a modified version of Hong and White (1995) to avoid a non-zero centering term. Let $\{(y_i, x_i)\}_{i=1}^n$ be an i.i.d. sample with $x_i \in \mathbb{R}^G$. We test

$$H_0 : \mathbb{E}[y \mid x] = \gamma_0' x \quad \text{a.s.}$$

against general alternatives. Let $b(x)$ be a K -dimensional vector of basis functions, and define

$$B = (b(x_1), \dots, b(x_n))', \quad P = B(B'B)^{-1}B', \quad M = I - P, \quad \hat{u}_i = y_i - \hat{\gamma}' x_i,$$

where $\hat{\gamma}$ is the OLS estimator. The test statistic of Y. Sun and Li (2006) is

$$T = \frac{K^{-1/2} \sum_{i=1}^n \sum_{j \neq i} P_{ij} \hat{u}_i \hat{u}_j}{\sqrt{\hat{V}}}, \quad \hat{V} = \frac{2}{K} \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 \hat{u}_i^2 \hat{u}_j^2.$$

Replacing the plug-in variance estimator by its cross-fit analogue yields

$$\hat{V}_{\text{cf}} = \frac{2}{K} \sum_{i=1}^n \sum_{j \neq i} \frac{P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2} \{\hat{u}_i(m'_i \hat{u})\} \{\hat{u}_j(m'_j \hat{u})\},$$

where m_i denotes the i th column of M . The specification-testing problem fits the framework of Section 2 with

$$a_{ij} = K^{-1/2} P_{ij}, \quad c_{ij} = \frac{2}{K} \frac{P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2}, \quad \ell_i = m_i.$$

Decompose the residual to locate the drift. Let $\theta_0(x) = \mathbb{E}[y \mid x]$, let $\alpha_* = \mathbb{E}[b(x)b(x)']^{-1} \mathbb{E}[b(x)y]$ be the coefficient of the best sieve approximation to it, and let γ_* be the probability limit of $\hat{\gamma}$. Then

$$\hat{u}_i = \underbrace{\alpha'_* b(x_i) - \gamma'_* x_i}_{\Delta_i} + \underbrace{y_i - \theta_0(x_i)}_{e_i} + \underbrace{\theta_0(x_i) - \alpha'_* b(x_i)}_{\omega_i} + x'_i \hat{\delta}, \quad \hat{\delta} = \gamma_* - \hat{\gamma},$$

which telescopes to $\hat{u}_i = y_i - \hat{\gamma}' x_i$ and is (6) with $H_i = x_i$.

Proposition 1. *Suppose Assumption 2 holds. Then the specification-testing problem satisfies the conditions of Theorems 1 and 2. Hence the cross-fit variance estimator is consistent for the asymptotic variance under the null hypothesis, and, under Assumption 2(iv), under the alternative as well, while the plug-in estimator is not.*

3.1.1. *Decomposition of $m'_i \hat{u}$.* Applying m'_i to $\hat{u}_i$ gives the second factor of $\hat{u}_i(m'_i \hat{u})$,

$$m'_i \hat{u} = m'_i \Delta + m'_i \omega + m'_i e + m'_i H \hat{\delta},$$

where $m'_i e$ is mean zero. We take the other three terms in turn.

The first term is the drift, and it is annihilated exactly: $m'_i \Delta = 0$ for every i . Because the basis spans the linear terms, both $B\alpha_*$ and $X\gamma_*$ lie in $\text{col}(B)$, so $\Delta \in \text{col}(B)$ and $M\Delta = (I - P)\Delta = 0$.

The second term is $m'_i \omega$, which is what the auxiliary factor leaves of the sieve approximation error. The systematic component is $\pi_i = \Delta_i + \omega_i = \theta_0(x_i) - \gamma'_* x_i$, the nonlinearity that the linear specification cannot represent; the sieve coefficient cancels, so π does not depend on the basis. Applying the auxiliary factor to it leaves $m'_i \pi = \lambda_i$, since $m'_i \Delta = 0$, and Assumption 2(iii) makes $\max_i |\omega_i|$ and $\max_i |\lambda_i|$, with $\lambda_i = m'_i \omega$, of order $o_p(b_n^{-3})$. The plug-in therefore squares π_i , which is of order one under the alternative, whereas the second factor retains only λ_i of it.

The last term vanishes as well: the auxiliary vectors annihilate the estimation directions. Because the basis spans the linear terms, $X = B\alpha_K$ for some α_K , and hence $\ell'_i H = m'_i X = 0$. By Lemma 1 this makes $g_{2i} = 0$ and $\mathcal{R}_n = 0$, so only $\mathcal{G}_1$ and $\mathcal{G}_2$ survive in (9).

3.2. Overidentification testing. We next consider the overidentification test for linear instrumental variable models studied by Chao *et al.* (2014). Let

$$y_i = \beta' x_i + u_i, \quad x_i = \Upsilon_i + v_i,$$

for $i = 1, \dots, n$, where $x_i \in \mathbb{R}^G$ is endogenous and $z_i \in \mathbb{R}^K$ is a vector of instruments with $K > G$. Following Chao *et al.* (2014), the instruments $Z = (z_1, \dots, z_n)'$ and the reduced form $\Upsilon = (\Upsilon_1, \dots, \Upsilon_n)'$ are treated as nonrandom, with $\mathbb{E}[u_i] = 0$ and $\mathbb{E}[v_i] = 0$, and Υ is spanned by the instruments, $\Upsilon = Z\Psi$.³ Let

$$P = Z(Z'Z)^{-1}Z', \quad M = I - P, \quad \hat{u}_i = y_i - \hat{\beta}' x_i,$$

where $\hat{\beta}$ has probability limit β_* , equal to β under the null. The choice of $\hat{\beta}$ can be the JIV estimators analyzed by Chao *et al.* (2012) or the HLIM and HFUL estimators of Hausman *et al.* (2012), for example. Chao *et al.* (2014) use the last one. Assumption 3(iii) asks no more of it than $\hat{\beta} \xrightarrow{P} \beta_*$.

Because Z is nonrandom and $\mathbb{E}[u_i] = 0$, the moment condition $\mathbb{E}[Z'u] = 0$ holds by construction, so it is not what the test can check. Write

$$\bar{u}_i = \mathbb{E}[y_i - \beta'_* x_i], \quad \bar{u} = (\bar{u}_1, \dots, \bar{u}_n)',$$

for the expected residual at the pseudo-true value. The null of correct specification is $H_0 : \bar{u} = 0$, tested against $H_1 : \bar{u} \neq 0$, which arises when the instruments affect the outcome directly or the reduced form is not explained by the structural equation. With $K > G$ the instruments impose $K - G$ overidentifying restrictions. The test statistic of Chao *et al.* (2014) is

$$J = \frac{\sum_{i=1}^n \sum_{j \neq i} P_{ij} \hat{u}_i \hat{u}_j}{\sqrt{\hat{\Phi}}} + K, \quad \hat{\Phi} = \frac{1}{K} \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 \hat{u}_i^2 \hat{u}_j^2.$$

The null hypothesis is rejected when J exceeds the $(1 - \alpha)$ quantile of the χ^2_{K-G} distribution.

Replacing the plug-in variance estimator by its cross-fit analogue yields

$$\hat{\Phi}_{\text{cf}} = \frac{1}{K} \sum_{i=1}^n \sum_{j \neq i} \frac{P_{ij}^2}{M_{ii} M_{jj} + M_{ij}^2} \{\hat{u}_i(m'_i \hat{u})\} \{\hat{u}_j(m'_j \hat{u})\},$$

³Chao *et al.* (2014) also allow the reduced form to be approximated rather than spanned, $\Upsilon_i = f_0(w_i)$ with $n^{-1} \sum_i \|\Upsilon_i - \Psi' z_i\|^2 \rightarrow 0$ for $z_i = (p_{1K}(w_i), \dots, p_{KK}(w_i))'$. We impose the exact case in Assumption 3(ii), because the mean-square condition does not bound $\max_i \|m'_i \Upsilon\|$, which the remainder term requires here.

where m_i denotes the i th column of M . The overidentification-testing problem fits the framework of Section 2 with

$$a_{ij} = P_{ij}, \quad c_{ij} = \frac{2P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2}, \quad \ell_i = m_i.$$

With these weights the variance estimators of Section 2 are exactly $2K$ times those in the statistic, $\hat{V} = 2K\hat{\Phi}$ and $\hat{V}_{\text{cf}} = 2K\hat{\Phi}_{\text{cf}}$. Because the conclusions of Section 2 are statements about ratios, they apply to $\hat{\Phi}$ and $\hat{\Phi}_{\text{cf}}$ unchanged.

Decompose the residual to locate the drift. Since $\hat{u}_i = (y_i - \beta'_*x_i) + x'_i(\beta_* - \hat{\beta})$ and $y_i - \beta'_*x_i = \bar{u}_i + \{(y_i - \beta'_*x_i) - \bar{u}_i\}$, splitting $\bar{u} = P\bar{u} + M\bar{u}$ gives

$$\hat{u}_i = \Delta_i + \underbrace{(y_i - \beta'_*x_i) - \bar{u}_i}_{e_i} + \underbrace{m'_i\bar{u} + x'_i(\beta_* - \hat{\beta})}_{\rho_i}, \quad \Delta = P\bar{u}.$$

This is (2) and (6) with $\omega_i = m'_i\bar{u}$, $H_i = x_i$ and $\hat{\delta} = \beta_* - \hat{\beta}$. The error e_i has mean zero by construction and equals $u_i + v'_i(\beta - \beta_*)$ whenever the structural equation $y_i = \beta'x_i + u_i$ holds. Under H_0 , we have $e_i = u_i$.

Proposition 2. *Suppose Assumption 3 holds. Then the overidentification-testing problem satisfies the conditions of Theorems 1 and 2. Hence the cross-fit variance estimator is consistent for the asymptotic variance under the null hypothesis, and, under Assumption 3(iv), under the alternative as well, while the plug-in estimator is not.*

The decomposition of $m'_i\hat{u}$ is as in Section 3.1. The drift is annihilated because $\Delta = P\bar{u}$ lies in $\text{col}(Z)$ and $M = I - P$ projects onto its orthogonal complement. The other two terms behave differently.

Here $\omega = M\bar{u}$ is the part of the systematic residual not spanned by the instruments, the overidentification analogue of the sieve error in the specification test. Because M is idempotent, $m'_i\omega = m'_i\bar{u} = \omega_i$: the auxiliary factor removes nothing from ω . In the notation of Section 2 this is $\lambda_i = \omega_i$, so the two maximal bounds on ω in Assumption 1 coincide. The systematic component is $\pi_i = \Delta_i + \omega_i = \bar{u}_i$, the expected residual itself, and since the drift is annihilated, the auxiliary factor leaves $m'_i\pi = \omega_i$. The plug-in therefore squares $\bar{u}_i$, whereas the second factor retains only ω_i .

The estimation directions are not annihilated, unlike in the specification test. Since Υ is spanned by the instruments but v is not,

$$\ell'_i H = m'_i x = m'_i v \neq 0,$$

so $g_{2i} \neq 0$ in Lemma 1 and the remainder $\mathcal{R}_n$ has to be bounded.

Remark 4. Since $m'_i \omega$ does not vanish, Assumption 3(iii) has to bound $\max_i |\omega_i|$ directly, and we describe which departures from the null satisfy that bound. Write the departure as a term entering the outcome equation directly, $y_i = \beta' x_i + h_i + u_i$. Then $\bar{u} = \Upsilon(\beta - \beta_*) + h$, and $M\Upsilon = 0$ because $\Upsilon = Z\Psi$, so $\omega = Mh$ whatever the pseudo-true value. Thus, Assumption 3(iii) asks that the part of the departure the instruments do not span be uniformly small, $\max_i |(Mh)_i| = o_p(b_n^{-3})$. Two situations justify the condition. First, h may lie in $\text{col}(Z)$, so that the instruments enter the outcome equation linearly and $Mh = 0$. Second, h may be a smooth function of the variables from which the instruments are built, as in the series construction of footnote 3, in which case Mh is an approximation error of the same kind as the sieve error of Section 3.1. The alternative against which Chao *et al.* (2014) report the power of their test is of the first kind, a direct effect of one instrument on the outcome with the coefficient on it moved away from zero, and the design of Section 4.2 follows that design. In contrast, Assumption 3(iii) rules out a case where h_i is driven by a variable the instruments do not span; in this case, the systematic component $\pi_i = \Delta_i + \omega_i$ enters the second factor of $\hat{u}_i(m'_i \hat{u})$ as well as the first. ■

3.3. Testing many linear restrictions. We finally consider testing many linear restrictions in linear regression models, where variance inflation under the alternative becomes particularly important as the number of restrictions increases. Let $\{(y_i, x_i)\}_{i=1}^n$ be an i.i.d. sample with $x_i = (w'_i, z'_i)'$, where $w_i \in \mathbb{R}^{d-r}$ is fixed-dimensional, with $d = \dim(x_i)$ the total number of regressors, and $z_i \in \mathbb{R}^r$, with the number of restrictions r potentially large. Consider the heteroskedastic linear regression model

$$y_i = w'_i \beta_1 + z'_i \beta_2 + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i \mid x_i] = 0,$$

and test $H_0 : \beta_2 = \beta_{20}$. Let $X = [W \ Z]$ and

$$M_W = I - W(W'W)^{-1}W', \quad Q = M_W Z(Z'M_W Z)^{-1}Z'M_W, \quad M = I - X(X'X)^{-1}X'.$$

Write Q_{ij} and M_{ij} for the (i, j) th entries of Q and M , respectively, and let m_i denote the i th column of M . Let $\hat{\sigma}_\varepsilon^2 = \frac{1}{n-d} \sum_{i=1}^n (y_i - x'_i \hat{\beta})^2$ denote the unrestricted residual variance estimator, where $\hat{\beta}$ is the unrestricted OLS estimator. The conventional F statistic is

$$F = \frac{(y - Z\beta_{20})' Q (y - Z\beta_{20})}{r \hat{\sigma}_\varepsilon^2}.$$

This is the statistic Anatolyev and Solvsten (2023) use for $H_0 : R\beta = q$, taken at $R = [0 \ I_r]$ and $q = \beta_{20}$; a general full-row-rank R can be brought to $[0 \ I_r]$ by a nonsingular transformation of the regressors, which leaves F unchanged.

Following Anatolyev and Solvsten (2023), consider critical values of the form

$$c_\alpha(E, V) = \frac{1}{r\hat{\sigma}_\varepsilon^2}(E + V^{1/2}\kappa_\alpha), \quad \kappa_\alpha = \frac{q_{1-\alpha}(F_{r,n-d}) - 1}{\sqrt{2/r + 2/(n-d)}},$$

and $q_{1-\alpha}(F_{r,n-d})$ denotes the $(1 - \alpha)$ quantile of the $F_{r,n-d}$ distribution. The null hypothesis is rejected whenever $F > c_\alpha(E, V)$. Anatolyev and Solvsten (2023) estimate E using a leave-one-out estimator and V using a leave-three-out estimator.

Anatolyev and Solvsten (2023, Remark 4) suggest estimating the error variances under the null, and write out the resulting leave-out estimates in their Appendix A.4. We follow that line in a narrower setting, where imposing the restriction leaves only a fixed-dimensional parameter to estimate, and we use the restricted residual itself in place of their leave-out construction. Consider the plug-in estimators

$$\hat{E} = \sum_{i=1}^n Q_{ii}\tilde{e}_i^2, \quad \hat{V}^{\text{plug}} = 2 \sum_{i=1}^n \sum_{j \neq i} Q_{ij}^2 \tilde{e}_i^2 \tilde{e}_j^2,$$

where $\tilde{e} = y - W\tilde{\beta}_1 - Z\beta_{20}$ is the restricted residual and

$$\tilde{\beta}_1 = (W'W)^{-1}W'(y - Z\beta_{20})$$

is the restricted OLS estimator under H_0 . Keeping the centering estimator unchanged, we replace the plug-in variance estimator by

$$\hat{V}^{\text{cf}} = 2 \sum_{i=1}^n \sum_{j \neq i} \frac{Q_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2} (\tilde{e}_i m'_i \tilde{e}) (\tilde{e}_j m'_j \tilde{e}).$$

The switch costs nothing in the statistic itself: $QW = 0$, so $(y - Z\beta_{20})'Q(y - Z\beta_{20}) = \tilde{e}'Q\tilde{e}$ and the numerator of F is unchanged.

Subtracting the centering estimator removes the diagonal terms, $\tilde{e}'Q\tilde{e} - \hat{E} = \sum_{i=1}^n \sum_{j \neq i} Q_{ij}\tilde{e}_i\tilde{e}_j$, so the centered statistic is exactly (1) with $\hat{u}_i = \tilde{e}_i$. The many-restrictions testing problem therefore fits the framework of Section 2 with

$$a_{ij} = Q_{ij}, \quad c_{ij} = 2 \frac{Q_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2}, \quad \ell_i = m_i.$$

Decompose the residual to locate the drift. By the definition of $\tilde{\beta}_1$, the restricted residual is $\tilde{e} = M_W(y - Z\beta_{20})$. Substituting $y_i = w'_i\beta_1 + z'_i\beta_2 + \varepsilon_i$ gives $y - Z\beta_{20} = W\beta_1 + \Delta_0 + \varepsilon$ with $\Delta_0 = Z(\beta_2 - \beta_{20})$, and $M_W W = 0$ leaves the exact identity $\tilde{e} = M_W(\Delta_0 + \varepsilon)$. Splitting it as

$$\hat{u}_i = \tilde{e}_i = \underbrace{(M_W \Delta_0)_i}_{\Delta_i} + \underbrace{\varepsilon_i}_{e_i} + \underbrace{-w'_i(W'W)^{-1}W'\varepsilon}_{\rho_i}$$

gives (2) and (6) with $\omega_i = 0$, $H_i = w_i$ and $\hat{\delta} = -(W'W)^{-1}W'\varepsilon$. Here $\hat{\delta} = \mathbb{E}[\tilde{\beta}_1 | X] - \tilde{\beta}_1$ is the sampling noise in the restricted estimator, free of the drift. The bound $\|\hat{\delta}\| = O_p(n^{-1/2})$, which Appendix D derives from Assumption 4(i), therefore holds under the alternative as well as the null, with the rate coming from the fixed dimension of w_i .

Proposition 3. *Suppose Assumption 4 holds. Then the many-restrictions testing problem satisfies the conditions of Theorems 1 and 2. Hence the cross-fit variance estimator is consistent for the asymptotic variance under the null hypothesis, and, under Assumption 4(iv), under the alternative as well, while the plug-in estimator is not.*

The decomposition of $m_i'\hat{u}$ is as in Section 3.1, with $\omega \equiv 0$ here, so that $\pi_i = \Delta_i$ and $\lambda_i = 0$. The drift is annihilated because $MW = 0$ gives $MP_W = 0$, and $\Delta_0 \in \text{col}(Z) \subseteq \text{col}(X)$ gives $M\Delta_0 = 0$; hence $M\Delta = MM_W\Delta_0 = 0$. The auxiliary vectors annihilate the estimation directions as well, $\ell_i'H = m_i'W = 0$, so $g_{2i} = 0$ and $\mathcal{R}_n = 0$ in Lemma 1, as in the specification test.

Remark 5. We close Section 3 by comparing the two biases across the three applications. Lemma 2 gives them in each case on substituting a_{ij} and $\ell_i = m_i$ from Table 1. The plug-in bias (10) behaves the same way in all three. It is positive, and under condition (iv) of the corresponding application-specific assumption it does not vanish relative to the variance: $\liminf_n B_n/V > 0$. The cross-fit bias differs across the three, according to whether λ vanishes.

In the specification test λ is negligible, and in the many-restrictions test $\lambda \equiv 0$, so in both the cross-fit bias reduces to the three terms of Remark 1. Each carries an off-diagonal entry of M or $M\Sigma M$. In the specification test that factor is bounded by the maximal leverage $\max_i P_{ii}$, which Assumption 2(ii) makes $o_p(1)$; in the many-restrictions test it is small once the off-diagonal entries of Q are not concentrated on a few pairs, which Assumption 4(iv) requires. Because the number of contaminating terms grows with the number of restrictions while the drift per observation stays bounded, the reduction matters most there.

In the overidentification test $\lambda_i = \omega_i$ does not vanish and the reduction fails. The terms of (11) carrying $\pi_i\lambda_i = \bar{u}_i\omega_i$ survive, and the leading three,

$$\sum_{i=1}^n \sum_{j \neq i} c_{ij} \{ \bar{u}_i\omega_i \bar{u}_j\omega_j + \bar{u}_i\omega_i M_{jj}\sigma_j^2 + M_{ii}\sigma_i^2 \bar{u}_j\omega_j \},$$

carry no off-diagonal factor: at most a diagonal entry of M , which is bounded away from zero. Accordingly $\max_i |\omega_i|$ has to be bounded directly there, and the gain in that application depends on how much of the misspecification the instruments span. ■

4. SIMULATION RESULTS

This section reports the finite-sample behavior of the cross-fit variance estimator in the three problems of Section 3. In every design the test statistic is held fixed and the procedures differ only in the variance estimator, or, in Section 4.3, in the critical value built from it, so the differences reported below are attributable to variance estimation alone. Rejection frequencies are at the nominal 5% level, and the null rejection rates are reported with each design. They are not size-corrected except in Appendix E, where each procedure is instead referred to the 95th percentile of its own simulated null distribution.

4.1. Specification testing. We first consider the specification-testing problem of Section 3.1. Following Tripathi and Kitamura (2003),

$$y_i = \beta_0 + \beta_1 x_i + \frac{c}{\sqrt{2\pi}} \exp(-(x_i - 2.5)^2) + u_i, \quad u_i = \varepsilon_i \sqrt{0.1 + 0.2x_i + 0.3x_i^2},$$

with $\beta_0 = \beta_1 = 1$, $n = 400$, $\log x_i \sim N(0, 1)$ truncated at the upper 5% tail, and $\varepsilon_i \sim N(0, 1)$ independent of x_i . The parameter c indexes the departure from linearity: $c = 0$ is the null, under which the linear specification is correct, and $c > 0$ the alternative, under which it is not.

The sieve is a cubic spline,

$$b(x) = (1, x, x^2, x^3, (x - \tau_1)_+^3, \dots, (x - \tau_{K-4})_+^3)',$$

with the knots τ_k at the $k/(K - 3)$ sample quantiles of x . The constant and the linear term belong to it, so it spans the linear specification, as Section 3.1 assumes. At a given noncentrality the inflation of the plug-in variance is of order $\sqrt{K}/n$ relative to the variance. We therefore vary K , reporting $K \in \{8, 20, 40\}$, which spans K/n from 0.02 to 0.10.

Table 2 reports the frequencies. Under the null the conventional test is mildly conservative at every K , while the cross-fit test is close to nominal at $K = 8$ and mildly liberal at $K = 40$, reaching 6.8%; the two effects act in the same direction, so part of the difference reported next is a difference in null rejection rates rather than in power. Under the alternative the cross-fit procedure dominates at every c and every K , and the margin widens with K/n : the largest gap is 2.4 percentage points at $K = 8$, 3.5 at $K = 20$ and 4.6 at $K = 40$. Within each K it is largest at intermediate c , where the contamination is big enough to cost power while both procedures still have room to gain.

TABLE 2. Empirical rejection frequencies for the specification test

c	$K = 8$		$K = 20$		$K = 40$	
	T	T_{cf}	T	T_{cf}	T	T_{cf}
0.0	3.5	5.2	3.6	6.1	4.5	6.8
1.5	14.2	16.6	12.2	15.2	9.9	13.7
2.0	25.0	26.6	20.0	23.5	16.4	20.7
2.5	39.0	41.2	32.2	35.7	25.7	30.3
3.0	55.5	57.1	46.8	49.9	37.8	42.3
3.5	71.1	72.0	62.0	64.8	51.7	55.4
4.0	82.9	83.1	76.7	77.8	65.9	69.0

Notes: Rejection frequencies in percent at the nominal 5% level, $n = 400, 5,000$ Monte Carlo replications. $c = 0$ is the null; $c > 0$ the alternative. T is the test of Y. Sun and Li (2006); T_{cf} replaces $\hat{V}$ by $\hat{V}_{\text{cf}}$. The cross-fit variance estimate is negative in at most 0.8% of replications; those replications are counted as non-rejections.

4.2. Overidentification testing. We next consider the overidentification test of Section 3.2. The model is

$$y_i = \beta' x_i + h_i + \varepsilon_i, \quad x_i = (1, \Upsilon_i + v_i)', \quad \Upsilon_i = z_i' \psi, \quad \varepsilon_i = \rho v_i + \sqrt{1 - \rho^2} s_i e_i,$$

with $\beta = 0$, $G = 2$, $\rho = 0.6$, $s_i^2 = (1 + z_{i1}^2)/2$, $v_i, e_i \sim N(0, 1)$ and $n = 400$, so that the conditional variance is driven by the first instrument. The instruments are $z_i = (1, z_{i1}, \dots, z_{i,K-1})'$, where z_{i2} is a dummy equal to one with probability 0.05 and the rest are standard normal; ψ is spread evenly over the non-constant instruments with the concentration parameter held at $n\psi'\psi = 400$, so identification does not weaken as K grows. We use the HFUL estimator of Hausman *et al.* (2012) as Chao *et al.* (2014) do.

The misspecification is $h_i = \eta z_{i3}$, a direct effect of one instrument on the outcome. It lies in $\text{col}(Z)$, so $\omega = M\bar{u} = 0$ and the whole of $\bar{u}$ enters the statistic as drift, $\Delta = P\bar{u} = \bar{u}$.

Table 3 reports the frequencies. Both statistics are close to nominal under the null, with the cross-fit version at most one percentage point above the conventional one. The cross-fit statistic has rejection frequencies at least as large at every η , and the margin grows with K/n : 0.5 percentage points at 0.025, 1.9 at 0.10 and 3.9 at 0.30.

4.3. Testing many linear restrictions. We finally consider the many-restrictions testing problem of Section 3.3. We draw $n = 300$ independent observations from

$$y_i = w_i' \beta_1 + z_i' \beta_2 + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | x_i] = 0, \quad \mathbb{E}[\varepsilon_i^2 | x_i] = (1 + 5|z_{i,1}|)^2,$$

without an intercept, where $x_i = (w_i', z_i')' \in \mathbb{R}^{1+r}$ is standard normal, w_i is scalar and $\beta_1 = 1$. The conditional variance is driven by the first restricted regressor.

TABLE 3. Empirical rejection frequencies for the overidentification test

η	$K = 10$		$K = 40$		$K = 120$	
	$K/n = 0.025$		$K/n = 0.10$		$K/n = 0.30$	
	J	J_{cf}	J	J_{cf}	J	J_{cf}
0.00	5.1	5.4	5.6	6.0	4.9	5.9
0.05	8.9	9.1	7.0	7.6	5.5	6.5
0.10	21.3	21.7	11.4	12.8	7.2	8.7
0.15	46.2	46.7	23.8	25.4	11.4	13.6
0.20	75.2	75.3	44.2	46.1	19.9	23.4
0.30	99.4	99.4	90.0	90.8	51.7	55.6

Notes: Rejection frequencies in percent at the nominal 5% level, $n = 400, 2,000$ Monte Carlo replications, first step estimated by HFUL. $\eta = 0$ is the null. J is the statistic of Chao *et al.* (2014); J_{cf} replaces $\hat{\Phi}$ by $\hat{\Phi}_{\text{cf}}$. The cross-fit variance estimate is positive in every replication.

We test $H_0 : \beta_2 = 0$ against $\beta_2 = (\eta/\sqrt{r})\mathbf{1}_r$, where $\eta \geq 0$ indexes the signal strength and $\eta = 0$ is the null. The alternative is composite, any $\beta_2 \neq 0$, and the departure enters the statistic through the quadratic form $\sum_{i=1}^n \sum_{j \neq i} Q_{ij} \Delta_i \Delta_j$, which is unchanged when β_2 is replaced by $-\beta_2$: the rejection frequency depends on the magnitude of β_2 and on its direction in $\mathbb{R}^r$, not on its sign. Taking β_2 proportional to $\mathbf{1}_r$ fixes that direction at the one spreading the departure evenly over the r coefficients, and η varies only the magnitude. The normalization by $\sqrt{r}$ holds the magnitude of the coefficient vector fixed as r varies, so the three values $r \in \{60, 120, 180\}$ are comparable and span r/n from 0.20 to 0.60. With standard normal regressors this design has $\max_i |\Delta_i|$ growing like $\sqrt{\log n}$, so it meets the boundedness condition of Assumption 4(iv) only up to a logarithmic factor; the calibrated design of Section 5.3, where the covariate is truncated, meets it exactly. Four procedures share the same F statistic and differ only in the critical value: the conventional F value, the procedure of Anatolyev and Sølvsten (2023), the plug-in procedure based on restricted residuals, and the proposed cross-fit procedure.

Figure 1 reports the frequencies. Under the null the cross-fit and Anatolyev–Sølvsten procedures are close to the nominal level at all three r , the plug-in procedure is conservative and becomes more so as r grows, from 3.8% at $r = 60$ to 2.8% at $r = 180$, and the conventional F test is mildly oversized. Under the alternative the cross-fit procedure rejects more often than the plug-in one at every η , and the margin widens sharply with r/n : the largest gap is 4.1 percentage points at $r = 60$, 7.9 at $r = 120$ and 13.6 at $r = 180$. The cross-fit procedure also tracks the Anatolyev–Sølvsten procedure closely throughout, exceeding it by at most 2.5 percentage points, at $r = 180$. In a setup where the null restriction leaves only a fixed-dimensional parameter to estimate, the cross-fit procedure avoids the leave-three-out construction of Anatolyev and Sølvsten (2023) but still achieves moderately higher

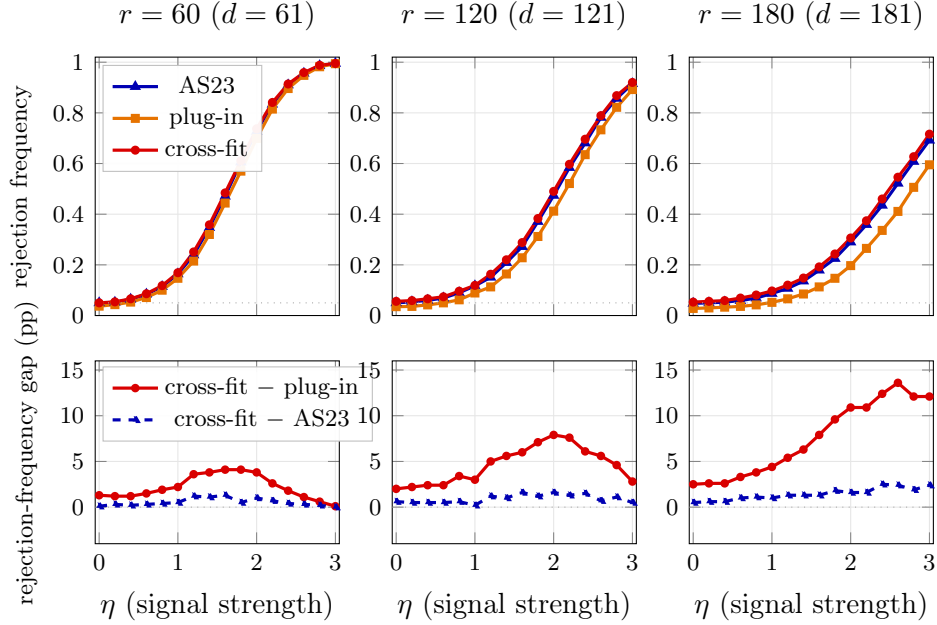


FIGURE 1. Empirical rejection frequencies under $\beta_2 = (\eta/\sqrt{r})\mathbf{1}_r$ for $r \in \{60, 120, 180\}$, based on $n = 300$ and 2,000 Monte Carlo replications. The regressors are independent standard normal and $\sigma_i^2 = (1 + 5|z_{i,1}|)^2$. The top row reports the rejection frequencies of AS23, the plug-in procedure, and the proposed cross-fit procedure. The bottom row reports the corresponding rejection-frequency gaps between the cross-fit procedure and the other two procedures. At $\eta = 0$, the rejection frequencies of the conventional F /AS23/plug-in/cross-fit procedures are, respectively, 6.1/5.0/3.8/5.0% for $r = 60$, 5.9/5.0/3.5/5.5% for $r = 120$, and 5.3/4.8/2.8/5.3% for $r = 180$.

power than theirs. In the OHIE specifications of Section 5, the leave-three-out procedure is computationally prohibitive and cannot be reported, whereas the cross-fit critical value is readily computed.

5. EMPIRICAL APPLICATION: OHIE JOINT HTE TEST

We illustrate the many-restrictions procedures on data from the Oregon Health Insurance Experiment (OHIE), a lottery-based expansion of access to Medicaid in Oregon in 2008. Finkelstein *et al.* (2012), hereafter F12, estimate intent-to-treat (ITT) effects of lottery selection on a broad range of health-care utilization and health outcomes, and report treatment-effect estimates for several prespecified subgroups while emphasizing the limited precision of those comparisons. We complement their analysis with a formal joint test of heterogeneity in the ITT effect: we interact lottery selection with a sieve expansion of the baseline covariates and test whether all interaction coefficients are jointly zero. Such a test

differs from inspecting a collection of individual subgroup coefficients because it accounts explicitly for the dependence among a potentially large number of restrictions.

Rejection would indicate that the data are inconsistent with a constant conditional ITT effect relative to the specified sieve alternative. Non-rejection means that the data do not detect such heterogeneity; it does not establish exact homogeneity outside the alternative class, nor rule out economically meaningful heterogeneity that is estimated imprecisely.

5.1. Setup. For each observation i , we estimate the linear probability model

$$y_i = w_i' \beta_1 + z_i' \beta_2 + \varepsilon_i, \quad H_0 : \beta_2 = 0, \quad (13)$$

where y_i indicates whether individual i had at least one non-childbirth overnight hospital admission during the preceding six months. The vector w_i contains a constant, the lottery-selection indicator D_i , the baseline covariates and corresponding missing-value indicators, and survey-wave-by-household-size fixed effects. The vector $z_i = D_i b(X_i)$ contains interactions between lottery selection and a sieve expansion $b(X_i)$ of the baseline covariates.

We use lottery-list participants who completed both the baseline survey and the 12-month survey and for whom the hospital-admission outcome is available, yielding $n = 16,455$ observations. Missing continuous baseline covariates are replaced by their sample medians and missing binary covariates by their sample modes. Corresponding missing-value indicators are included among the controls.⁴

The sieve contains a rich polynomial expansion of the baseline covariates. It includes main effects, polynomials in age and household income up to degree five, all pairwise interactions among the binary covariates, and interactions between the binary covariates and the polynomial terms in the continuous covariates. The full-sample specification contains $r = 136$ restrictions.

Following F12, we also conduct separate analyses for individuals reporting fair or poor baseline health and for those reporting good baseline health. Within each subgroup, sieve terms that become collinear with the controls are removed, leaving $r = 112$ restrictions.⁵

We compare the conventional F critical value, the plug-in critical value based on restricted residuals, and the proposed cross-fit critical value. The leave-three-out procedure of Anatolyev and Slvsten (2023) is not reported because its computational cost is prohibitive for the full OHIE specifications considered here.

⁴This imputation scheme is used only to retain observations in the joint regression. The resulting test should therefore be interpreted relative to the baseline variables and missing-value controls included in the specified model.

⁵The subgroup samples exclude observations that do not belong to the corresponding self-rated-health category. Their sample sizes therefore need not sum to the full-sample size.

TABLE 4. OHIE joint tests of heterogeneity in the ITT effect

Configuration	n	r	r/n	F	c_α^F	c_α^{plug}	c_α^{cf}	Rejection
Full sample	16,455	136	0.83%	1.009	1.209	1.091	1.090	None
Fair/poor baseline health	6,365	112	1.76%	0.839	1.588	1.032	1.031	None
Good baseline health	9,610	112	1.17%	0.880	1.588	1.105	1.104	None

Notes: c_α^F , c_α^{plug} , and c_α^{cf} denote the 5% critical values based on the conventional F approximation, the restricted-residual plug-in variance estimator, and the proposed cross-fit variance estimator, respectively. The procedure of Anatolyev and Sølvsten (2023) is not reported because of its computational cost in these specifications.

5.2. Results. Table 4 reports the results. None of the three procedures rejects the null hypothesis at the 5% level, in the full sample or in either self-rated-health subgroup.

The plug-in and cross-fit critical values are nearly identical in this application. This is consistent with a setting in which the realized drift contamination is small at the estimated specification. Because the empirical result is a single realization, however, the similarity of the two critical values does not by itself reveal how the procedures behave under alternatives.

5.3. Calibrated simulation. We next consider a simulation design motivated by the empirical application. The purpose is not to reproduce the OHIE data exactly, but to study the relative performance of the competing procedures in a design with a binary outcome, treatment interactions, a moderately large sieve, and heterogeneous treatment effects concentrated in the tails of a baseline covariate. We simulate from (13) with $\mathbb{E}[\varepsilon_i | x_i] = 0$ and test the same null, $H_0 : \beta_2 = 0$.

Each observation contains a binary treatment indicator D_i and baseline covariates

$$X_i = (\text{age}_i, \text{income}_i, \text{female}_i, \text{white}_i, \text{urban}_i)'$$

The continuous covariates age_i and income_i are generated independently from $N(0, 0.25)$ truncated to $[-2, 2]$. The binary covariates are generated independently from Bernoulli distributions with success probabilities 0.55, 0.70, and 0.75, respectively. We set $w_i = (1, D_i, X_i')'$ and $z_i = D_i b(X_i)$, where $b(X_i)$ is a polynomial sieve containing main effects, higher-order polynomial terms, and interactions. The coefficients in β_1 are calibrated to approximately match the magnitudes observed in the empirical application.

Under the alternative, the conditional mean contains the heterogeneity term

$$\Delta_{0i} = z_i' \beta_2 = \eta D_i \phi(X_i),$$

so that the drift of Section 3.3 is $\Delta = M_W \Delta_0$, where $\eta \geq 0$ controls the magnitude of treatment-effect heterogeneity. We consider $\phi(X_i) = \text{age}_i^3$ and $\phi(X_i) = \text{age}_i^5$. The null hypothesis corresponds to $\eta = 0$. These alternatives concentrate heterogeneous treatment

TABLE 5. Null rejection frequencies as r/n increases

Design	r/n	Conventional F	AS23	Plug-in	Cross-fit
$\text{age}^3, n = 1,500$	1.8%	0.080	0.042	0.040	0.042
$\text{age}^3, n = 1,000$	2.7%	0.080	0.050	0.048	0.052
$\text{age}^3, n = 750$	3.6%	0.110	0.046	0.044	0.056
$\text{age}^3, n = 500$	5.4%	0.110	0.038	0.040	0.050
$\text{age}^5, n = 1,500$	1.8%	0.088	0.036	0.038	0.040
$\text{age}^5, n = 1,000$	2.8%	0.092	0.032	0.032	0.038
$\text{age}^5, n = 750$	3.7%	0.100	0.036	0.038	0.038
$\text{age}^5, n = 500$	5.6%	0.124	0.048	0.046	0.062

Notes: Entries are rejection frequencies under $\eta = 0$ at the nominal 5% significance level, based on 500 Monte Carlo replications.

effects in the tails of the age distribution, with the fifth-order alternative generating a more uneven drift pattern.

The baseline simulation uses $n = 1,500$, with $r = 27$ restrictions under the cubic alternative and $r = 28$ under the fifth-order one.

We report rejection frequencies at the nominal 5% level. At the baseline sample size the Anatolyev–Sølvsten, plug-in and cross-fit procedures reject between 3.6% and 4.4% under the null, while the conventional F procedure rejects between 8.0% and 8.8%. Holding the sieve dimension fixed and reducing n to 1,000, 750 and 500 raises r/n from 1.8% to 5.6%, and Table 5 shows that the pattern persists: the F test is oversized throughout, from 8.0% to 12.4%, while the other three stay close to nominal, the cross-fit procedure becoming mildly liberal at the smaller sample sizes and reaching 6.2% under the fifth-order basis at $n = 500$.

Figure 2 reports raw rejection frequencies under the alternatives. The cross-fit procedure rejects more often than the plug-in procedure at every point of the grid. The largest gap is 5.6 percentage points under the cubic alternative and 7.6 under the fifth-order one, where it stays between 4 and 8 points over the intermediate range $\eta \in [0.075, 0.2]$; against the Anatolyev–Sølvsten procedure the largest gap is about 4 points. The gaps close as η grows and all three procedures approach rejection probability one.

The gain depends on how the drift is distributed. It is larger under the tail-concentrated fifth-order alternative than under the cubic one throughout, and it persists as r/n grows. Appendix E reports size-corrected frequencies across the four sample sizes, where the cubic gap falls from 5.0 percentage points at $r/n = 1.8\%$ to about zero at 5.4% while the fifth-order gap stays between 4.0 and 7.0 points. This is consistent with the mechanism of Section 2, in which the gain is governed by B_n/V and a concentrated systematic component makes that ratio large. The comparison at the largest r/n should be read with care, since the asymptotic approximation behind the cross-fit procedure becomes less accurate there.

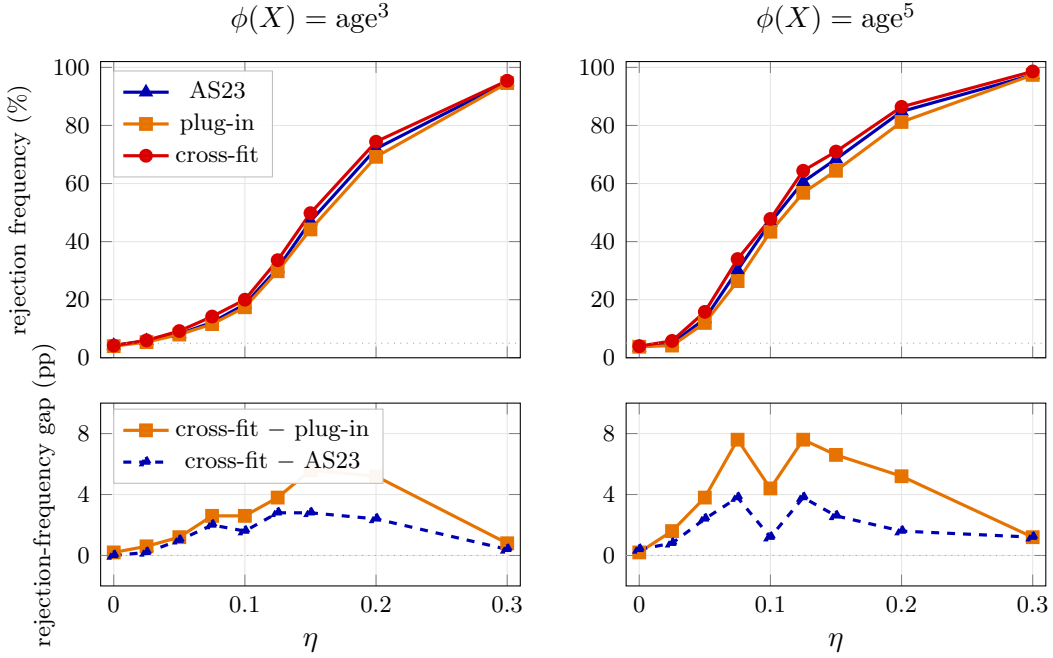


FIGURE 2. Raw rejection frequencies in the OHIE-calibrated simulation. The columns correspond to $\phi(X) = \text{age}^3$ ($r = 27$) and $\phi(X) = \text{age}^5$ ($r = 28$). The sample size is $n = 1,500$, and results are based on 500 Monte Carlo replications for each value of η . The top row reports raw rejection frequencies for the Anatolyev–Sølvsten, plug-in, and cross-fit procedures. The bottom row reports the differences between the cross-fit rejection frequency and those of the other two procedures. Rejection frequencies are computed over replications in which the corresponding variance estimate is nonnegative. Across the 4,500 replications for each alternative direction, the Anatolyev–Sølvsten variance estimate is negative in 29 replications under the cubic alternative and 137 replications under the fifth-order alternative. The cross-fit variance estimate is never negative across the 9,000 replications.

5.3.1. *Numerical stability.* The simulations also reveal a substantial difference in numerical stability. Across the two baseline designs, the proposed cross-fit variance estimate is never negative in 9,000 replications. The Anatolyev–Sølvsten variance estimate is negative in 29 replications under the cubic alternative and 137 replications under the fifth-order alternative, with most occurrences arising at larger values of η . The same holds in the smaller samples of Appendix E: it is negative in as many as 50 of 500 replications at a given grid point, whereas the cross-fit estimate is never negative in any reported design.

These results provide finite-sample evidence that the proposed cross-fit estimator is numerically more stable in the designs considered here. They should not, however, be interpreted as implying that the cross-fit variance estimator is algebraically guaranteed to be nonnegative.

Although $M_{ii}M_{jj} + M_{ij}^2 > 0$ under a full-rank design, the products entering the numerator of the cross-fit estimator can have either sign.

APPENDIX A. PROOFS FOR SECTION 2

A.1. Proof of Lemma 1. By (2) and (6), $\hat{u}_i = \check{u}_i + H'_i \hat{\delta}$ and $\ell'_i \hat{u} = (\ell'_i \check{u}) + h_i \hat{\delta}$. Multiplying the two factors gives

$$\hat{u}_i(\ell'_i \hat{u}) = \check{u}_i(\ell'_i \check{u}) + \{\check{u}_i h_i + (\ell'_i \check{u}) H'_i\} \hat{\delta} + (\hat{\delta}' H_i)(h_i \hat{\delta}).$$

The coefficient in braces is a $1 \times p$ row vector whose transpose is g_{1i} . In the last term both factors are scalars, so it equals $\hat{\delta}'(H_i h_i) \hat{\delta}$; since $\hat{\delta}' A \hat{\delta} = \hat{\delta}'(A + A') \hat{\delta} / 2$ for any $p \times p$ matrix A , and $(H_i h_i)' = h'_i H'_i$, it equals $\hat{\delta}' g_{2i} \hat{\delta}$. This proves the first assertion.

Substituting this identity into (5) and collecting terms by their degree in $\hat{\delta}$ gives (7), with $\mathcal{G}_1$ and $\mathcal{G}_2$ as stated and with

$$\mathcal{R}_n = \sum_{i=1}^n \sum_{j \neq i} c_{ij} \{ (g'_{1i} \hat{\delta})(\hat{\delta}' g_{2j} \hat{\delta}) + (g'_{1j} \hat{\delta})(\hat{\delta}' g_{2i} \hat{\delta}) + (\hat{\delta}' g_{2i} \hat{\delta})(\hat{\delta}' g_{2j} \hat{\delta}) \}.$$

Each term of the last display is homogeneous in $\hat{\delta}$: the first two are $\|\hat{\delta}\|^3$ times the corresponding term at $\hat{\delta}/\|\hat{\delta}\|$, and the third is $\|\hat{\delta}\|^4$ times it. Since $\hat{\delta}/\|\hat{\delta}\|$ is a unit vector, taking absolute values and then the supremum over unit vectors gives the stated bound.

Finally, if $\ell'_i H = 0$ for every i then $h_i = 0$, so $g_{2i} = 0$ and $g_{1i} = (\ell'_i \check{u}) H_i$. Every term in $\mathcal{R}_n$ carries a factor g_{2i} or g_{2j} , so $\mathcal{R}_n = 0$.

A.2. Proof of Theorem 1. Under H_0 we have $\Delta_i = 0$, so $\check{u}_i = e_i + \omega_i$ in the notation of Lemma 1, and (7) gives

$$\hat{V}_{\text{cf}} = \sum_{i=1}^n \sum_{j \neq i} c_{ij} g_{0i} g_{0j} + \mathcal{G}'_1 \hat{\delta} + \hat{\delta}' \mathcal{G}_2 \hat{\delta} + \mathcal{R}_n.$$

The last three terms are $o_p(V)$: by the Cauchy–Schwarz inequality $|\mathcal{G}'_1 \hat{\delta}| \leq \|\hat{\delta}\| \|\mathcal{G}_1\|$ and $|\hat{\delta}' \mathcal{G}_2 \hat{\delta}| \leq \|\hat{\delta}\|^2 \|\mathcal{G}_2\|$, and (9) bounds the first two. For the third, Lemma 1 gives $|\mathcal{R}_n| \leq \|\hat{\delta}\|^3 \mathcal{G}_3 + \|\hat{\delta}\|^4 \mathcal{G}_4$, which is $o_p(V)$ by (9). It therefore suffices to show that $\sum_{i=1}^n \sum_{j \neq i} c_{ij} g_{0i} g_{0j} = V + o_p(V)$.

Write $g_{0i} = e_i(\ell'_i e) + R_i$ with

$$R_i = \omega_i(\ell'_i e) + e_i(\ell'_i \omega) + \omega_i(\ell'_i \omega).$$

By Assumption 1, $\max_i |e_i(\ell'_i e)| = O_p(b_n^2)$ and

$$\begin{aligned} \max_i |R_i| &\leq \max_i |\omega_i| \max_i |\ell'_i e| + \max_i |e_i| \max_i |\ell'_i \omega| + \max_i |\omega_i| \max_i |\ell'_i \omega| \\ &= o_p(b_n^{-2}), \end{aligned}$$

so that $\max_{i,j} \{|e_i(\ell'_i e) R_j| + |R_i e_j(\ell'_j e)| + |R_i R_j|\} = o_p(1)$. By (3),

$$\begin{aligned} \sum_{i=1}^n \sum_{j \neq i} c_{ij} g_{0i} g_{0j} - V &= \sum_{i=1}^n \sum_{j \neq i} \{c_{ij} e_i(\ell'_i e) e_j(\ell'_j e) - 2a_{ij}^2 \mathbb{E}[e_i^2] \mathbb{E}[e_j^2]\} \\ &\quad + \sum_{i=1}^n \sum_{j \neq i} c_{ij} (e_i(\ell'_i e) R_j + R_i e_j(\ell'_j e) + R_i R_j). \end{aligned}$$

The first term is $o_p(V)$ by (8). For the second, Assumption 1 gives $C_n = O(A_n)$, so

$$\begin{aligned} \left| \sum_{i=1}^n \sum_{j \neq i} c_{ij} (e_i(\ell'_i e) R_j + R_i e_j(\ell'_j e) + R_i R_j) \right| &\leq C_n \max_{i,j} \{|e_i(\ell'_i e) R_j| + |R_i e_j(\ell'_j e)| + |R_i R_j|\} \\ &= O(A_n) o_p(1) = o_p(A_n), \end{aligned}$$

which is $o_p(V)$ because $V \geq 2c^2 A_n$. The conclusion follows.

A.3. Proof of Lemma 2. Under H_1 we have $\tilde{u}_i = \pi_i + e_i$ with $\pi_i = \Delta_i + \omega_i$, so the leading term of Lemma 1 is $g_{0i} = \tilde{u}_i(\ell'_i \tilde{u}) = \check{A}_i$ and $\sum_{i=1}^n \sum_{j \neq i} c_{ij} g_{0i} g_{0j} = \check{V}_{\text{cf}}$.

For the plug-in estimator, $\tilde{u}_i^2 = \pi_i^2 + 2\pi_i e_i + e_i^2$ gives $\mathbb{E}[\tilde{u}_i^2] = \pi_i^2 + \sigma_i^2$, and for $i \neq j$ the independence of e_i and e_j gives $\mathbb{E}[\tilde{u}_i^2 \tilde{u}_j^2] = \mathbb{E}[\tilde{u}_i^2] \mathbb{E}[\tilde{u}_j^2]$. Subtracting $V = 2 \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \sigma_i^2 \sigma_j^2$ yields (10).

For the cross-fit estimator, since $\ell'_i \Delta = 0$, we have $\ell'_i \tilde{u} = \lambda_i + \ell'_i e$ with $\lambda_i = \ell'_i \omega$, so

$$\check{A}_i = \pi_i \lambda_i + \pi_i (\ell'_i e) + \lambda_i e_i + e_i (\ell'_i e).$$

Multiplying $\check{A}_i$ by $\check{A}_j$ for $i \neq j$ gives sixteen products, which we take in order of their degree in e , counting $\ell'_i e$ as degree one. The one of degree zero is $\pi_i \lambda_i \pi_j \lambda_j$. The four of degree one carry a single mean-zero factor and vanish. Of the six of degree two, $\lambda_i \lambda_j \mathbb{E}[e_i e_j] = 0$, and the rest follow from

$$\mathbb{E}[(\ell'_i e)(\ell'_j e)] = (L \Sigma L')_{ij}, \quad \mathbb{E}[(\ell'_i e) e_j] = L_{ij} \sigma_j^2, \quad \text{and} \quad \mathbb{E}[e_i (\ell'_i e)] = L_{ii} \sigma_i^2.$$

A product of degree three survives only if its three error factors can share one index. In $\lambda_i e_i \cdot e_j (\ell'_j e)$ two of them are already fixed at i and j , so it vanishes, and the same holds for $e_i (\ell'_i e) \cdot \lambda_j e_j$; in $\pi_i (\ell'_i e) \cdot e_j (\ell'_j e)$ only e_j is fixed and the two free indices can both be set to j , giving $\mathbb{E}[(\ell'_i e) e_j (\ell'_j e)] = L_{ij} L_{jj} \mu_{3j}$, and symmetrically $\mathbb{E}[e_i (\ell'_i e) (\ell'_j e)] = L_{ji} L_{ii} \mu_{3i}$. The single product of degree four is

$$\mathbb{E}[e_i (\ell'_i e) e_j (\ell'_j e)] = (L_{ii} L_{jj} + L_{ij} L_{ji}) \sigma_i^2 \sigma_j^2,$$

where only second moments enter because $i \neq j$. Collecting the nine surviving products, $\mathbb{E}[\check{A}_i \check{A}_j]$ is the sum of the eight terms in braces in (11) and $(L_{ii} L_{jj} + L_{ij} L_{ji}) \sigma_i^2 \sigma_j^2$. Multiplying

by c_{ij} and summing over $i \neq j$ gives $\mathbb{E}[\check{V}_{\text{cf}}]$, in which the last term contributes exactly V by the definition of c_{ij} ; subtracting V leaves (11).

A.4. Proof of Theorem 2. Decompose

$$\hat{V}_{\text{cf}} - V = (\hat{V}_{\text{cf}} - \check{V}_{\text{cf}}) + (\check{V}_{\text{cf}} - \mathbb{E}[\check{V}_{\text{cf}}]) + B_n^{\text{cf}}.$$

The first term is $o_p(V)$ by the third condition of (12), the second is $o_p(V)$ by the first, and the third is $o(V)$ by the last. Hence $\hat{V}_{\text{cf}}/V \xrightarrow{p} 1$.

Similarly

$$\hat{V} - (V + B_n) = (\hat{V} - \check{V}) + (\check{V} - \mathbb{E}[\check{V}]),$$

both terms being $o_p(V + B_n)$ by the fourth and second conditions of (12). Hence $\hat{V}/(V + B_n) \xrightarrow{p} 1$.

For the last assertion, every summand of (10) is nonnegative, so discarding the quartic term and using $\sigma_i^2 \geq c$ together with the symmetry of the weights,

$$B_n \geq 2 \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 (\pi_i^2 \sigma_j^2 + \sigma_i^2 \pi_j^2) \geq 4c \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \pi_i^2.$$

Since $V \leq 2C^2 A_n$,

$$\frac{B_n}{V} \geq \frac{2c}{C^2} \cdot \frac{1}{A_n} \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \pi_i^2,$$

which is bounded away from zero by assumption. The cross-fit bias (11) admits no matching lower bound: its terms carry $\pi_i \lambda_i$ and μ_{3i} , whose signs are unrestricted.

A.5. **Proof of Corollary 1.** By Theorem 2, $\hat{V}_{\text{cf}}/V \xrightarrow{p} 1$ and $\hat{V}/(V + B_n) \xrightarrow{p} 1$. Writing

$$\frac{\hat{V}}{\hat{V}_{\text{cf}}} = \frac{\hat{V}}{V + B_n} \cdot \frac{V + B_n}{V} \cdot \frac{V}{\hat{V}_{\text{cf}}}$$

and rearranging gives $\hat{V}/\hat{V}_{\text{cf}} = (1 + B_n/V)(1 + o_p(1))$. The two statistics share the numerator Q_n , so $T_n^{\text{cf}}/T_n = (\hat{V}/\hat{V}_{\text{cf}})^{1/2}$, and taking square roots gives the first assertion. If $\liminf_n B_n/V \geq \gamma > 0$ then $P(\hat{V} \geq \hat{V}_{\text{cf}}) \rightarrow 1$, and on that event $|T_n^{\text{cf}}| = |Q_n|/\sqrt{\hat{V}_{\text{cf}}} \geq |Q_n|/\sqrt{\hat{V}} = |T_n|$, whatever the sign of Q_n . A two-sided test rejecting when $|T_n|$ exceeds a common critical value therefore rejects no less often when $\hat{V}_{\text{cf}}$ is used.

A.6. **Proof of Corollary 2.** By Theorem 2, $\hat{V} = (V + B_n)\{1 + o_p(1)\}$, and by assumption $V + B_n = S_n\{1 + o(1)\}$, so $\hat{V} = S_n\{1 + o_p(1)\}$. With $Q_n = N_n + o_p(S_n^{1/2})$,

$$T_n = \frac{Q_n}{\hat{V}^{1/2}} = \frac{N_n + o_p(S_n^{1/2})}{S_n^{1/2}\{1 + o_p(1)\}} = \{\tau_n + o_p(1)\}\{1 + o_p(1)\},$$

which is the first assertion. For the second, Corollary 1 gives $T_n^{\text{cf}}/T_n = (1 + B_n/V)^{1/2}\{1 + o_p(1)\}$. Every term of (10) is nonnegative, so $B_n \geq S_n$, and $V + B_n = S_n\{1 + o(1)\}$ forces $V/S_n \rightarrow 0$, hence $B_n/V \rightarrow \infty$ and the ratio diverges.

Finally, suppose $\limsup_n |\tau_n| < c$. The first assertion gives $|T_n| \leq \{|\tau_n| + o_p(1)\}\{1 + o_p(1)\}$, so $P(|T_n| > c) \rightarrow 0$. If in addition $\liminf_n |\tau_n| > 0$, then $|T_n|$ is bounded away from zero with probability approaching one, and $|T_n^{\text{cf}}| = |T_n|(T_n^{\text{cf}}/T_n) \xrightarrow{p} \infty$, so $P(|T_n^{\text{cf}}| > c) \rightarrow 1$.

A.7. Sufficient conditions for Theorem 2. The first two conditions of (12) and the requirement $B_n^{\text{cf}} = o(V)$ are statements about weighted sums over pairs of observations. This appendix reduces them to a few aggregate quantities that each application can bound directly, which keeps the verifications below short. Recall $\pi_i = \Delta_i + \omega_i$, $\lambda_i = \ell'_i \omega$, and write

$$\xi_n = \max_i \max(|\pi_i|, |\lambda_i|), \quad C_2 = \sum_{i=1}^n \sum_{j \neq i} c_{ij}^2, \quad \bar{c}_j = \sum_{i \neq j} |c_{ij}|,$$

$$L_F^2 = \sum_{i=1}^n \sum_{j \neq i} L_{ij}^2, \quad S_F^2 = \sum_{i=1}^n \sum_{j \neq i} (L \Sigma L')_{ij}^2.$$

Let $\tilde{C}$ be the matrix with off-diagonal entries c_{ij} and zero diagonal, and $D_v = \text{diag}(v)$. Note that $\sum_j \bar{c}_j = C_n$ using the notation of Section 2.

Lemma 3. *Suppose Assumption 1 holds, that $\ell'_i \Delta = 0$, $\|L\| \leq 1$, $\max_i \mathbb{E}[e_i^6] \leq C$, and that (8) holds. If $\xi_n = O(1)$ and*

$$C_2 = o(A_n^2), \quad \max_j \bar{c}_j = o(A_n), \quad C_2^{1/2}(L_F + S_F) = o(A_n), \quad (\text{A.1})$$

then the first two conditions of (12) hold and $B_n^{\text{cf}} = o(V)$.

Proof. The proof has five steps. Steps 1–3 give the first condition of (12), Step 4 the second, and Step 5 the bound $B_n^{\text{cf}} = o(V)$. We use repeatedly that $\text{Var}(e'Be) \leq C\|B\|_F^2$ for symmetric B (for general B , apply it to $(B + B')/2$, whose Frobenius norm is no larger) and independent mean-zero errors with four bounded moments, and that $\|B_1 B_2\|_F \leq \min(\|B_1\| \|B_2\|_F, \|B_1\|_F \|B_2\|)$. Since $\|L\| \leq 1$ we have $(LL')_{jj} \leq 1$ for every j .

Step 1. Decomposition into T_1 and T_2 . Put $\nu_i = e_i(\ell'_i e) - L_{ii}\sigma_i^2$, $\alpha_i = \mathbb{E}[\check{A}_i] = \pi_i \lambda_i + L_{ii}\sigma_i^2$ and $\eta_i = \check{A}_i - \alpha_i = \pi_i(\ell'_i e) + \lambda_i e_i + \nu_i$, and let $\bar{\alpha}_j = \sum_{i \neq j} c_{ij} \alpha_i$. Because $c_{ij} = c_{ji}$,

$$\sum_{i=1}^n \sum_{j \neq i} c_{ij} \{\check{A}_i \check{A}_j - \mathbb{E}[\check{A}_i \check{A}_j]\} = 2 \sum_j \bar{\alpha}_j \eta_j + \sum_{i=1}^n \sum_{j \neq i} c_{ij} \{\eta_i \eta_j - \mathbb{E}[\eta_i \eta_j]\} =: T_1 + T_2.$$

Step 2. The term T_1 . We show that $T_1 = o_p(V)$. Since $V \geq 2c^2A_n$, it is sufficient by Chebyshev's inequality to show $\text{Var}(T_1) = o(A_n^2)$, which we now derive from the hypothesis $\max_j \bar{c}_j = o(A_n)$. Substituting $\eta_j = \pi_j(\ell'_j e) + \lambda_j e_j + \nu_j$ into $T_1 = 2 \sum_j \bar{\alpha}_j \eta_j$ yields

$$\frac{1}{2}T_1 = \underbrace{\sum_j \bar{\alpha}_j \pi_j(\ell'_j e)}_{(L'g)'e} + \underbrace{\sum_j \bar{\alpha}_j \lambda_j e_j}_{(D_{\bar{\alpha}}\lambda)'e} + \underbrace{\sum_j \bar{\alpha}_j \nu_j}_{e'D_{\bar{\alpha}}Le - \mathbb{E}[e'D_{\bar{\alpha}}Le]},$$

where $g_j = \bar{\alpha}_j \pi_j$ and $D_{\bar{\alpha}} = \text{diag}(\bar{\alpha}_1, \dots, \bar{\alpha}_n)$. The first identity holds because $\ell'_j e$ is the j th entry of Le , and the third because the constants inside the ν_j sum to $\sum_j \bar{\alpha}_j L_{jj} \sigma_j^2 = \mathbb{E}[e'D_{\bar{\alpha}}Le]$. Collecting the first two gives

$$\frac{1}{2}T_1 = h'e + \{e'Ne - \mathbb{E}[e'Ne]\}, \quad h := L'g + D_{\bar{\alpha}}\lambda, \quad N := D_{\bar{\alpha}}L.$$

The linear form has variance at most $C\|h\|^2$. Since $e'Ne = e'N_s e$ with $N_s = (N + N')/2$ symmetric, the quadratic form has variance at most $C\|N_s\|_F^2$. Since $\text{Var}(X + Y) \leq 2\text{Var} X + 2\text{Var} Y$, it follows that $\text{Var}(T_1) \leq C\{\|h\|^2 + \|N_s\|_F^2\}$. Now

$$\|h\|^2 \leq 2\|g\|^2 + 2\sum_j \bar{\alpha}_j^2 \lambda_j^2 \leq C\sum_j \bar{\alpha}_j^2 (1 + \pi_j^2 + \lambda_j^2),$$

where the first inequality follows from $\|L'g\| \leq \|L\|\|g\|$ and $\|L\| \leq 1$, and the second from $\|g\|^2 = \sum_j \bar{\alpha}_j^2 \pi_j^2$. Moreover

$$\|N_s\|_F^2 \leq \|N\|_F^2 = \sum_j \bar{\alpha}_j^2 (LL')_{jj} \leq \sum_j \bar{\alpha}_j^2,$$

where the first inequality is the triangle inequality for $\|\cdot\|_F$ together with $\|N'\|_F = \|N\|_F$, and the last is $(LL')_{jj} \leq 1$. Since $\xi_n = O(1)$ and $\sigma_i^2 \leq C$ and $|L_{ii}| \leq \|L\| \leq 1$ we have $|\alpha_i| \leq C$, hence $|\bar{\alpha}_j| \leq C\bar{c}_j$ and

$$\sum_j \bar{\alpha}_j^2 (1 + \pi_j^2 + \lambda_j^2) \leq C\sum_j \bar{c}_j^2 \leq C\left(\max_j \bar{c}_j\right)C_n \leq C\left(\max_j \bar{c}_j\right)A_n = o(A_n^2),$$

where the second inequality follows from $\sum_j \bar{c}_j = C_n$, and the last follows from $C_n \leq (2/c_L)A_n$; the same bound covers $\|N_s\|_F^2$, so $\text{Var}(T_1) = o(A_n^2)$.

Step 3. The term T_2 . We show that $T_2 = o_p(V)$. Expanding $\eta_i = \pi_i(\ell'_i e) + \lambda_i e_i + \nu_i$ gives

$$\begin{aligned} \eta_i \eta_j &= \pi_i \pi_j (\ell'_i e)(\ell'_j e) + \lambda_i \lambda_j e_i e_j + \pi_i \lambda_j (\ell'_i e) e_j + \lambda_i \pi_j e_i (\ell'_j e) \\ &\quad + \pi_i (\ell'_i e) \nu_j + \pi_j (\ell'_j e) \nu_i + \lambda_i e_i \nu_j + \lambda_j e_j \nu_i + \nu_i \nu_j, \end{aligned}$$

which splits T_2 into nine groups, four quadratic in e , four cubic and one quartic. The quartic group is covered by the hypothesis (8); each of the other eight is shown to have variance $o(A_n^2)$, which with $V \geq 2c^2 A_n$ and Chebyshev's inequality makes it $o_p(V)$.

The quartic group $\sum_{i=1}^n \sum_{j \neq i} c_{ij} \{\nu_i \nu_j - \mathbb{E}[\nu_i \nu_j]\}$ involves neither π nor λ . Substituting $e_i(\ell'_i e) = \nu_i + L_{ii} \sigma_i^2$ and using $c_{ij} = c_{ji}$,

$$\sum_{i=1}^n \sum_{j \neq i} c_{ij} e_i(\ell'_i e) e_j(\ell'_j e) = \sum_{i=1}^n \sum_{j \neq i} c_{ij} \{\nu_i \nu_j - \mathbb{E}[\nu_i \nu_j]\} + 2 \sum_j \bar{w}_j \nu_j + \sum_{i=1}^n \sum_{j \neq i} c_{ij} (L_{ii} L_{jj} + L_{ij} L_{ji}) \sigma_i^2 \sigma_j^2.$$

Here $\bar{w}_j = \sum_{i \neq j} c_{ij} L_{ii} \sigma_i^2$, and the centering uses $\mathbb{E}[\nu_i \nu_j] = L_{ij} L_{ji} \sigma_i^2 \sigma_j^2$ for $i \neq j$. By the definition of c_{ij} the third sum equals $2 \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \sigma_i^2 \sigma_j^2 = V$, so subtracting V from both sides gives

$$\sum_{i=1}^n \sum_{j \neq i} \{c_{ij} e_i(\ell'_i e) e_j(\ell'_j e) - 2a_{ij}^2 \sigma_i^2 \sigma_j^2\} = \sum_{i=1}^n \sum_{j \neq i} c_{ij} \{\nu_i \nu_j - \mathbb{E}[\nu_i \nu_j]\} + 2 \sum_j \bar{w}_j \nu_j,$$

whose left-hand side is that of (8). As in the treatment of T_1 , $\sum_j \bar{w}_j \nu_j = e' D_{\bar{w}} L e - \mathbb{E}[e' D_{\bar{w}} L e]$, and $|L_{ii} \sigma_i^2| \leq C$ gives $|\bar{w}_j| \leq C \bar{c}_j$, so its variance is at most $C \sum_j \bar{w}_j^2 \leq C(\max_j \bar{c}_j) A_n = o(A_n^2)$. The quartic group is therefore $o_p(V)$ by (8).

Four groups are quadratic in e . The two carrying one π and one λ are equal by the symmetry of c , so the four give three centered quadratic forms,

$$\begin{aligned} \sum_{i=1}^n \sum_{j \neq i} c_{ij} \pi_i \pi_j \{(\ell'_i e)(\ell'_j e) - \mathbb{E}[(\ell'_i e)(\ell'_j e)]\} &= e' B_1 e - \mathbb{E}[e' B_1 e], & B_1 &:= L' D_{\pi} \tilde{C} D_{\pi} L, \\ \sum_{i=1}^n \sum_{j \neq i} c_{ij} \lambda_i \lambda_j \{e_i e_j - \mathbb{E}[e_i e_j]\} &= e' B_2 e - \mathbb{E}[e' B_2 e], & B_2 &:= D_{\lambda} \tilde{C} D_{\lambda}, \\ 2 \sum_{i=1}^n \sum_{j \neq i} c_{ij} \pi_i \lambda_j \{(\ell'_i e) e_j - \mathbb{E}[(\ell'_i e) e_j]\} &= 2\{e' B_3 e - \mathbb{E}[e' B_3 e]\}, & B_3 &:= L' D_{\pi} \tilde{C} D_{\lambda}. \end{aligned}$$

By $\|L\| \leq 1$, their Frobenius norms are at most $\|D_{\pi} \tilde{C} D_{\pi}\|_F$, $\|D_{\lambda} \tilde{C} D_{\lambda}\|_F$ and $\|D_{\pi} \tilde{C} D_{\lambda}\|_F$. Each of these three matrices has zero diagonal and (i, j) entry bounded in modulus by $\xi_n^2 |c_{ij}|$, so each norm is at most $\xi_n^2 C_2^{1/2}$ and each group has variance $O(\xi_n^4 C_2) = o(A_n^2)$.

The remaining four groups are cubic in e . They involve the products $\pi_i(\ell'_i e) \nu_j$, $\pi_j(\ell'_j e) \nu_i$, $\lambda_i e_i \nu_j$ and $\lambda_j e_j \nu_i$. Relabelling i and j and using $c_{ij} = c_{ji}$ turns the second group into the first and the fourth into the third, so only two distinct sums arise,

$$\sum_{i=1}^n \sum_{j \neq i} c_{ij} \pi_i(\ell'_i e) \nu_j \quad \text{and} \quad \sum_{i=1}^n \sum_{j \neq i} c_{ij} \lambda_i e_i \nu_j,$$

each appearing twice. By Lemma 4 below, the variance of each sum is at most $C\xi_n^2\mathcal{C}_2$: take $G = L$ and $v = \pi$ for the first, $G = I$ and $v = \lambda$ for the second. Since $\xi_n = O(1)$ and $\mathcal{C}_2 = o(A_n^2)$, both are $o(A_n^2)$. In the second, $U = \tilde{C}D_\lambda$ has zero diagonal, so the pieces bounded by $\sum_k U_{kk}^2$ vanish. This exhausts the nine terms in T_2 , so $T_2 = o_p(V)$, and with T_1 the first condition of (12) follows.

Step 4. The plug-in estimator. The second condition of (12) is the same statement for the plug-in estimator. The argument just given yields it at $L = I$, $\lambda = \pi$ and weights $2a_{ij}^2$ in place of c_{ij} . Since $\|L\| \leq 1$ gives $L_{ii}L_{jj} + L_{ij}L_{ji} \leq 2$ and hence $c_{ij} \geq a_{ij}^2$, the plug-in analogues of $\mathcal{C}_2$ and of $\max_j \bar{c}_j$ are bounded up to constants by the quantities that (A.1) controls. At $L = I$ the two Frobenius quantities vanish, $L_F = S_F = 0$. The target is $V + B_n$, which is no smaller than V , so a bound of $o(A_n^2)$ still suffices. The one group that does not transfer is the quartic one, since (8) is a hypothesis about the cross-fit weights. But at $L = I$ the variables $\nu_i = e_i^2 - \sigma_i^2$ are independent across i , so

$$\text{Var}\left(\sum_{i=1}^n \sum_{j \neq i} 2a_{ij}^2 \{\nu_i \nu_j - \mathbb{E}[\nu_i \nu_j]\}\right) = 2 \sum_{i=1}^n \sum_{j \neq i} (2a_{ij}^2)^2 \mathbb{E}[\nu_i^2] \mathbb{E}[\nu_j^2] \leq C\mathcal{C}_2 = o(A_n^2),$$

which needs only four moments.

Step 5. The cross-fit bias. It remains to show the third conclusion, $B_n^{\text{cf}} = o(V)$. We bound the eight terms of (11). Assumption 1 gives $\max_i |\lambda_i| = o_p(1)$, so the first three terms are at most $C\xi_n C_n (\max_i |\lambda_i|)(1 + \xi_n (\max_i |\lambda_i|)) = o(A_n)$. By the Cauchy–Schwarz inequality the fourth is at most

$$\left(\sum_{i=1}^n \sum_{j \neq i} c_{ij}^2 \pi_i^2 \pi_j^2\right)^{1/2} S_F \leq \xi_n^2 \mathcal{C}_2^{1/2} S_F,$$

the fifth and sixth are at most $C\xi_n^2 \mathcal{C}_2^{1/2} L_F$, and the last two are at most $C\xi_n \mathcal{C}_2^{1/2} L_F$, the third moments satisfying $|\mu_{3i}| \leq C$. All are $o(A_n)$ by (A.1), and $V \geq 2c^2 A_n$, so $B_n^{\text{cf}} = o(V)$. $\square$

Lemma 4. *Suppose Assumption 1 holds, that $\|L\| \leq 1$ and $\max_i \mathbb{E}[e_i^6] \leq C$, and put $\nu_i = e_i(\ell'_i e) - L_{ii}\sigma_i^2$. Let G be a matrix with rows g'_i and $\|G\| \leq 1$, and let v be a vector with $\max_i |v_i| \leq \xi_n$. Then*

$$\text{Var}\left(\sum_{i=1}^n \sum_{j \neq i} c_{ij} v_i (g'_i e) \nu_j\right) \leq C\xi_n^2 \mathcal{C}_2.$$

Proof. The argument has three parts. First, the sum is written as a cubic form in e plus a linear form. Second, the cubic form is split by index coincidence into seven pieces. Third, the variance of each piece is bounded by $\|\gamma\|_F^2$, by $\sum_k U_{kk}^2$ or directly by $\xi_n^2 \mathcal{C}_2$, the first two being themselves at most $\xi_n^2 \mathcal{C}_2$.

First, substituting $g'_i e = \sum_k G_{ik} e_k$ and then $\nu_j = e_j \sum_m L_{jm} e_m - L_{jj} \sigma_j^2$ gives

$$\begin{aligned} \sum_{i=1}^n \sum_{j \neq i} c_{ij} v_i (g'_i e) \nu_j &= \sum_j \sum_k \left(\sum_{i \neq j} c_{ij} v_i G_{ik} \right) e_k \nu_j \\ &= \sum_j \sum_k U_{jk} e_k \left(e_j \sum_m L_{jm} e_m - L_{jj} \sigma_j^2 \right) \\ &= \sum_{k,j,m} \gamma_{kjm} e_k e_j e_m - q' e, \end{aligned} \quad (\text{A.2})$$

where $U = \tilde{C} D_v G$. The second equality uses that its (j, k) entry is $\sum_{i \neq j} c_{ij} v_i G_{ik}$, where we use the symmetry of c and the zero diagonal of $\tilde{C}$. The third collects the cubic terms with $\gamma_{kjm} = L_{jm} U_{jk}$ and the linear ones with $q_k = \sum_j \gamma_{kjj} \sigma_j^2$; write $\|\gamma\|_F^2 = \sum_{k,j,m} \gamma_{kjm}^2$.

Second, the cubic form $\sum_{k,j,m} \gamma_{kjm} e_k e_j e_m - \sum_{k,j,m} \mathbb{E}[\gamma_{kjm} e_k e_j e_m]$ can be decomposed into

$$P_3 + Q_1 + Q_2 + Q_3 + \sum_k \gamma_{kkk} (e_k^3 - \mu_{3k}) + (\ell^{(1)} + \ell^{(2)} + \ell^{(3)})' e, \quad (\text{A.3})$$

where P_3 collects the terms with k, j, m distinct, the Q_r are the three two-index-coincidence sums

$$\begin{aligned} Q_1 &= \sum_{k=1}^n \sum_{m \neq k} \gamma_{kkm} (e_k^2 - \sigma_k^2) e_m, \\ Q_2 &= \sum_{k=1}^n \sum_{j \neq k} \gamma_{kjk} (e_k^2 - \sigma_k^2) e_j, \\ Q_3 &= \sum_{k=1}^n \sum_{j \neq k} \gamma_{kjj} (e_j^2 - \sigma_j^2) e_k, \end{aligned}$$

one for each of the three ways two of the indices of γ_{kjm} can coincide. In each of them the error whose index is repeated appears squared and centered, and the remaining error appears linearly. The three linear coefficients are $\ell_m^{(1)} = \sum_{k \neq m} \gamma_{kkm} \sigma_k^2$, $\ell_j^{(2)} = \sum_{k \neq j} \gamma_{kjk} \sigma_k^2$ and $\ell_k^{(3)} = \sum_{j \neq k} \gamma_{kjj} \sigma_j^2$.

Combining (A.2), (A.3) and $\mathbb{E}[q' e] = 0$ gives

$$\sum_{i=1}^n \sum_{j \neq i} c_{ij} v_i \{ (g'_i e) \nu_j - \mathbb{E}[(g'_i e) \nu_j] \} = P_3 + Q_1 + Q_2 + Q_3 + \sum_k \gamma_{kkk} (e_k^3 - \mu_{3k}) + (\ell^{(1)} + \ell^{(2)} + \ell^{(3)})' e - q' e.$$

Note that the two vectors $\ell_k^{(3)} = \sum_{j \neq k} \gamma_{kjj} \sigma_j^2$ and $q_k = \sum_j \gamma_{kjj} \sigma_j^2$ differ only in the summand $j = k$, so $\ell^{(3)} - q = -(\gamma_{kkk} \sigma_k^2)_k$. Then the last display is

$$P_3 + Q_1 + Q_2 + Q_3 + \sum_k \gamma_{kkk} (e_k^3 - \mu_{3k}) - \sum_k \gamma_{kkk} \sigma_k^2 e_k + (\ell^{(1)} + \ell^{(2)})' e. \quad (\text{A.4})$$

Finally, each piece of (A.4) has variance at most $C\xi_n^2\mathcal{C}_2$. The bounds below use two quantities, both at most $\xi_n^2\mathcal{C}_2$. The first is $\|\gamma\|_F^2$:

$$\|\gamma\|_F^2 = \sum_j (LL')_{jj} \|U_{j\cdot}\|^2 \leq \|U\|_F^2 \leq \|\tilde{C}D_v\|_F^2 \|G\|^2 \leq \sum_{i=1}^n \sum_{j \neq i} c_{ij}^2 v_i^2 \leq \xi_n^2 \mathcal{C}_2,$$

where the equality uses $\gamma_{kjm} = L_{jm}U_{jk}$ and $\sum_m L_{jm}^2 = (LL')_{jj}$, the first inequality $(LL')_{jj} \leq \|L\|^2 \leq 1$, the second $U = \tilde{C}D_v G$ together with $\|B_1 B_2\|_F \leq \|B_1\|_F \|B_2\|$, the third $\|G\| \leq 1$, and the last $v_i^2 \leq \xi_n^2$. The second quantity is $\sum_k U_{kk}^2$, for which $U_{kk} = \sum_{i \neq k} c_{ki} v_i G_{ik}$ and $(G'G)_{kk} \leq \|G\|^2 \leq 1$ give $\sum_k U_{kk}^2 \leq \xi_n^2 \mathcal{C}_2$ by the Cauchy–Schwarz inequality.

Now take the pieces in turn.

$\text{Var}(P_3)$: Expanding it pairs two coefficients γ_{kjm} and $\gamma_{k'j'm'}$, and the expectation vanishes unless (k', j', m') is a permutation of (k, j, m) , since the errors are independent with mean zero and the three indices of a triple are distinct. Each of the six permutations contributes at most $\|\gamma\|_F^2$ by the Cauchy–Schwarz inequality, so $\text{Var}(P_3) \leq C\|\gamma\|_F^2$.

$\text{Var}(Q_r)$: The same bound holds for all three, where the paired coefficients carry the same unordered pair of indices. For example,

$$\text{Var}(Q_1) = \sum_{k=1}^n \sum_{m \neq k} \gamma_{kkm}^2 (\mathbb{E}[e_k^4] - \sigma_k^4) \sigma_m^2 + \sum_{k=1}^n \sum_{m \neq k} \gamma_{kkm} \gamma_{mmk} \mu_{3k} \mu_{3m},$$

in which the second sum pairs γ_{kkm} with γ_{mmk} : the multisets differ, but the index pair $\{k, m\}$ is the same. Since $|\mu_{3k}| \leq (\mathbb{E}[e_k^4])^{3/4} \leq C$ and $2|\gamma_{kkm} \gamma_{mmk}| \leq \gamma_{kkm}^2 + \gamma_{mmk}^2$, that sum is at most $C\|\gamma\|_F^2$; it vanishes altogether when the errors are symmetric.

$\text{Var}(\sum_k \gamma_{kkk}(e_k^3 - \mu_{3k}))$: The diagonal piece has variance

$$\sum_k \gamma_{kkk}^2 \text{Var}(e_k^3) \leq C \sum_k U_{kk}^2,$$

and this is the only place the sixth moments are needed.

$\text{Var}(\sum_k \gamma_{kkk} \sigma_k^2 e_k)$ and $\text{Var}((\ell^{(1)} + \ell^{(2)})'e)$: A linear form $c'e$ has variance at most $C\|c\|^2$, so it is enough to bound $\sum_k \gamma_{kkk}^2 \sigma_k^4$, $\|\ell^{(1)}\|^2$ and $\|\ell^{(2)}\|^2$. The first is bounded by

$$\sum_k L_{kk}^2 U_{kk}^2 \sigma_k^4 \leq C \sum_k U_{kk}^2,$$

where the inequality follows from $|L_{kk}| \leq \|L\| \leq 1$ and $\sigma_k^2 \leq C$. For $\ell^{(1)}$, put $d_k = U_{kk}$, so that $\ell_m^{(1)} = \sum_{k \neq m} L_{km} U_{kk} \sigma_k^2 = (L' \Sigma d)_m - L_{mm} U_{mm} \sigma_m^2$. Then

$$\|\ell^{(1)}\|^2 \leq 2\|L' \Sigma d\|^2 + 2 \sum_m L_{mm}^2 U_{mm}^2 \sigma_m^4 \leq C \sum_k U_{kk}^2,$$

where the first inequality is $\|a - b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and the second follows from $\|L\| \leq 1$, $|L_{mm}| \leq \|L\| \leq 1$ and $\sigma_k^2 \leq C$. For $\ell^{(2)}$, the entries are $\ell_j^{(2)} = \sum_{i \neq j} c_{ij} v_i (L \Sigma G')_{ji} - \gamma_{jjj} \sigma_j^2$. Splitting the two parts as before and applying the Cauchy–Schwarz inequality to the first, $\sum_i (L \Sigma G')_{ji}^2 = (L \Sigma G' G \Sigma L')_{jj} \leq C^2$ gives $\|\ell^{(2)}\|^2 \leq C \xi_n^2 \mathcal{C}_2$. The seven pieces are dependent, so $\text{Var}(\sum_{r=1}^7 X_r) \leq 7 \sum_r \text{Var}(X_r)$ bounds the variance of the sum by $C \xi_n^2 \mathcal{C}_2$. $\square$

APPENDIX B. ASSUMPTIONS AND PROOF OF PROPOSITION 1

Let $\theta_0(x) = \mathbb{E}[y \mid x]$.

Assumption 2. *The following conditions hold, in the notation of Section 3.1.*

(i): $\{(y_i, x_i)\}_{i=1}^n$ are i.i.d. and $\mathbb{E}[e_i \mid x_i] = 0$ with $e_i = y_i - \theta_0(x_i)$. Under H_0 , $\theta_0(x) = \gamma'_0 x$ for some γ_0 . The conditional variances satisfy $0 < c \leq \sigma_i^2 := \mathbb{E}[e_i^2 \mid x_i] \leq C < \infty$ uniformly in i , and $\max_i \mathbb{E}[|e_i|^6 \mid x_i] \leq C$. The regressors have six unconditional moments, $\mathbb{E}\|x_i\|^6 < \infty$.

(ii): The basis dimension satisfies $K \rightarrow \infty$ and $K/n \rightarrow 0$. Let $B = (b(x_1), \dots, b(x_n))'$, $P = B(B'B)^{-1}B'$, and $M = I - P$. The basis is normalized so that $\mathbb{E}[b(x)b(x)'] = I_K$, and there is a sequence $\{\zeta_K\}$ with $\sup_x \|b(x)\| \leq \zeta_K$ and $\zeta_K^2 K/n \rightarrow 0$. In addition, the basis spans the linear terms: there exists a $K \times G$ matrix α_K with $b(x)' \alpha_K = x'$ almost surely. In particular $X = B \alpha_K$, so $MX = 0$.

(iii): There is a deterministic sequence $1 \leq b_n = o(n)$ such that, conditionally on $X = (x_1, \dots, x_n)'$,

$$\max_i |e_i| = O_p(b_n), \quad \max_i |m'_i e| = O_p(b_n), \quad \max_i |\omega_i| + \max_i |m'_i \omega| = o_p(b_n^{-3}), \quad \|\hat{\delta}\| = O_p(n^{-1/2}),$$

where m_i is the i th column of M .

(iv): Under the alternative the systematic component is bounded and non-degenerate:

$$\max_i |\pi_i| = O_p(1), \quad \liminf_n A_n^{-1} \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \pi_i^2 > 0,$$

where $\pi_i = \Delta_i + \omega_i = \theta_0(x_i) - \gamma'_* x_i$.

The normalization in (ii) is without loss of generality: P , M and $\hat{V}_{\text{cf}}$ are invariant to nonsingular linear transformations of b , so one may always pass to the transformed basis. The condition asks that the normalized basis satisfy the growth condition on ζ_K , as in Newey (1997, Assumption 3). In (iii) one may always take $b_n = n^{1/6}$, since $\mathbb{E}[\max_i |e_i|^6 \mid X] \leq \sum_i \mathbb{E}[|e_i|^6 \mid x_i] \leq Cn$ by (i), and $m'_i e$ has bounded sixth moments by Rosenthal's inequality; the restriction $b_n = o(n)$ therefore costs nothing, and the force of (iii) is the bound it places on ω . With $b_n = n^{1/6}$ that bound needs to be $\max_i |\omega_i| + \max_i |m'_i \omega| = o_p(n^{-1/2})$.

For splines and power series, the approximation condition in Newey (1997, Assumption 3) holds with rate $K^{-s/G}$ when the conditional expectation has s continuous derivatives, so the requirement is $\sqrt{n}K^{-s/G} \rightarrow 0$. The first condition of (iv) asks that the deviation of the conditional expectation from its best linear predictor be bounded, which is vacuous under the null, where the deviation is zero, and it enters through $\xi_n = O_p(1)$ in Lemma 3 and through the moment bounds on $\check{A}_i$ and $\check{u}_i$ in the estimation-error step under the alternative. The second makes the plug-in estimator inconsistent as required in Theorem 2.

Proof. For the specification testing problem,

$$a_{ij} = K^{-1/2}P_{ij}, \quad c_{ij} = \frac{2}{K} \frac{P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2}, \quad \ell_i = m_i,$$

where $M = I - P$ and m_i denotes the i -th column of M . We verify the conditions of Theorems 1 and 2. All arguments are conditional on $X = (x_1, \dots, x_n)'$. To simplify the presentation we give the argument for $G = 1$, that is, for scalar x ; the vector case only replaces the products $x_i x_j$ by $x_i x_j'$.

Step 1. The maintained assumption. We begin by verifying Assumption 1. By (ii), $\mathbb{E}\|n^{-1}B'B - I\|^2 \leq \zeta_K^2 K/n \rightarrow 0$, so $\lambda_{\min}(n^{-1}B'B) \xrightarrow{p} 1$ (Newey, 1997, proof of Theorem 1), and

$$P_{ii} \leq \lambda_{\min}^{-1}(n^{-1}B'B) n^{-1} \max_i \|b(x_i)\|^2 \leq \lambda_{\min}^{-1}(n^{-1}B'B) \zeta_K^2/n = o_p(1).$$

Two consequences are used below. Since P is symmetric and idempotent, $\sum_{i,j} P_{ij}^2 = \text{tr}(P) = K$, while $\sum_i P_{ii}^2 \leq (\max_i P_{ii})K$, so that

$$\sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 = K - \sum_i P_{ii}^2 = K + o_p(K);$$

and since $|P_{ij}| \leq \sqrt{P_{ii}P_{jj}}$ for a projection, also

$$\sum_{i=1}^n \sum_{j \neq i} P_{ij}^4 \leq \left(\max_i P_{ii}\right)^2 \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 = o_p(K) = o_p(K^2).$$

Consequently $M_{ii} = 1 - P_{ii}$ is bounded away from zero uniformly in i , say $\min_i M_{ii} \geq c_M > 0$, with probability approaching one. Since $M_{ij} = -P_{ij}$ for $i \neq j$, $M_{ii}M_{jj} + M_{ij}^2 \geq c_M^2$ uniformly over $i \neq j$ with probability approaching one. This is the leverage condition of Assumption 1 with $c_L = c_M^2$. First,

$$A_n = \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 = K^{-1} \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 = 1 + o_p(1).$$

Moreover,

$$2c^2 A_n \leq V = 2 \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \sigma_i^2 \sigma_j^2 \leq 2C^2 A_n,$$

because σ_i^2 is bounded above and away from zero. Also,

$$C_n = \sum_{i=1}^n \sum_{j \neq i} |c_{ij}| \leq \frac{C}{K} \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 = O_p(1) = O_p(A_n).$$

The maximal bounds for e_i , $m'_i e$, ω_i , and $m'_i \omega$ are imposed in Assumption 2(iii). This establishes Assumption 1.

Step 2. Orthogonality of the drift. We next verify the orthogonality condition in Theorem 2. Under the alternative, by the decomposition in Section 3.1, $\Delta = B\alpha_* - X\gamma_*$, which lies in $\text{col}(B)$ by Assumption 2(ii). As $M = I - P$ projects onto the orthogonal complement of $\text{col}(B)$, we have $M\Delta = 0$, hence $m'_i \Delta = 0$ for every i , and the systematic component is

$$\pi_i = \Delta_i + \omega_i = \theta_0(x_i) - \gamma'_* x_i, \quad \lambda_i = m'_i \omega.$$

Step 3. The null hypothesis. First we check condition (8). Under H_0 , $\Delta = 0$ and $\hat{u}_i = e_i + \rho_i$. Letting $B_i = e_i(m'_i e)$, it holds $\mathbb{E}[B_i B_j | X] = \mathbb{E}[e_i(m'_i e)e_j(m'_j e) | X]$ for $i \neq j$. Since

$$m'_i e = M_{ii} e_i + \sum_{k \neq i} M_{ki} e_k, \quad m'_j e = M_{jj} e_j + \sum_{\ell \neq j} M_{\ell j} e_\ell,$$

independence and mean-zero errors imply $\mathbb{E}[B_i B_j | X] = \sigma_i^2 \sigma_j^2 (M_{ii} M_{jj} + M_{ij}^2)$, and thus

$$\mathbb{E} \left[\sum_{i=1}^n \sum_{j \neq i} c_{ij} B_i B_j \middle| X \right] = \frac{2}{K} \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 \sigma_i^2 \sigma_j^2 = V.$$

Let $S_n = \sum_{i=1}^n \sum_{j \neq i} c_{ij} \{B_i B_j - \mathbb{E}[B_i B_j | X]\}$. Since $V \geq 2c^2 A_n = 2c^2 + o_p(1)$, it suffices to show $S_n = o_p(1)$. Writing $\tilde{P}_{ij}^2 = P_{ij}^2 / (M_{ii} M_{jj} + M_{ij}^2)$, we have $c_{ij} = \frac{2}{K} \tilde{P}_{ij}^2$ and $2a_{ij}^2 = \frac{2}{K} P_{ij}^2$, so that

$$\sum_{i=1}^n \sum_{j \neq i} \{c_{ij} B_i B_j - 2a_{ij}^2 \sigma_i^2 \sigma_j^2\} = \frac{2}{K} \sum_{i=1}^n \sum_{j \neq i} \{\tilde{P}_{ij}^2 e_i(m'_i e)e_j(m'_j e) - P_{ij}^2 \sigma_i^2 \sigma_j^2\}.$$

Apart from the factor $2/K$, this is the expression of Mikusheva and L. Sun (2022, Lemma 2) with P the sieve projection here, $M = I - P$ and $U_i^{(1)} = U_i^{(2)} = U_i^{(3)} = e_i$; their weight is $\tilde{P}_{ij}^2$ and their normalization is K^{-1} . Their conditions hold here: $K \rightarrow \infty$ by (ii), $P_{ii} \leq c_P < 1$ for some c_P because $\max_i P_{ii} = o_p(1)$ as shown above, and the errors are independent with mean zero and bounded sixth conditional moments by (i). The lemma therefore gives $S_n = o_p(1)$.

To see where the sixth moments of Assumption 2(i) enter, substitute $e_i(m'_i e) = \sum_k M_{ik} e_i e_k$, which makes S_n a centered quartic form in e ,

$$S_n = \sum_{i=1}^n \sum_{j \neq i} \sum_{k,l} \Gamma_{ijkl} \{e_i e_j e_k e_l - \mathbb{E}[e_i e_j e_k e_l | X]\}, \quad \Gamma_{ijkl} = c_{ij} M_{ik} M_{jl}.$$

Since $i \neq j$, no index occurs more than three times in a quartet, and hence no more than six times in $\mathbb{E}[S_n^2 | X]$; six is attained by the terms with $i' = i$ and $j' = j$, so no moment above order six enters. Since $V \geq 2c^2 A_n$, the conclusion $S_n = o_p(1)$ of the lemma gives

$$\sum_{i=1}^n \sum_{j \neq i} \{c_{ij} e_i(m'_i e) e_j(m'_j e) - 2a_{ij}^2 \sigma_i^2 \sigma_j^2\} = o_p(V).$$

This is (8).

Next we check condition (9). Under H_0 we have $\theta_0(x) = \gamma'_0 x$ by (i), so $\mathbb{E}[b(x)y] = \mathbb{E}[b(x)x']\gamma_0$; the representation $b(x)' \alpha_K = x'$ of (ii) turns the right-hand side into $\mathbb{E}[b(x)b(x)']\alpha_K \gamma_0$, so that $\alpha_* = \alpha_K \gamma_0$ and $\alpha'_* b(x) = \gamma'_0 x$. With $\hat{\gamma} \xrightarrow{p} \gamma_0$ this gives $\gamma_* = \gamma_0$, hence $\Delta_i = \omega_i = 0$, so $\tilde{u}_i = e_i$ in the notation of Lemma 1 and $\ell'_i H = m'_i X = 0$ by (ii). Then $g_{2i} = 0$ and $\mathcal{R}_n = 0$, while $g_{0i} = e_i(m'_i e)$ and $g_{1i} = (m'_i e)x_i$, so that, c_{ij} being symmetric,

$$\mathcal{G}_1 = 2 \sum_{i=1}^n \sum_{j \neq i} c_{ij} e_i(m'_i e) x_j(m'_j e), \quad \mathcal{G}_2 = \sum_{i=1}^n \sum_{j \neq i} c_{ij} x_i x_j(m'_i e)(m'_j e).$$

Since $\|\hat{\delta}\| = O_p(n^{-1/2})$ and $V \geq 2c^2 A_n = 2c^2 + o_p(1)$, it suffices to show $|\mathcal{G}_1| = o_p(n^{1/2})$ and $|\mathcal{G}_2| = o_p(n)$.

To see $|\mathcal{G}_1| = o_p(n^{1/2})$, let $\bar{c}_i = \sum_{j \neq i} |c_{ij}|$, so that $\sum_i \bar{c}_i = C_n$. Applying the Cauchy–Schwarz inequality over the pairs $i \neq j$ gives

$$\left| \sum_{i=1}^n \sum_{j \neq i} c_{ij} \alpha_i \beta_j \right| = \left| \sum_{i=1}^n \sum_{j \neq i} (|c_{ij}|^{1/2} \alpha_i) (|c_{ij}|^{1/2} \beta_j) \operatorname{sgn}(c_{ij}) \right| \leq \left(\sum_{i=1}^n \bar{c}_i \alpha_i^2 \right)^{1/2} \left(\sum_{j=1}^n \bar{c}_j \beta_j^2 \right)^{1/2} \quad (\text{B.1})$$

for any α and β . Take $\alpha_i = e_i(m'_i e)$ and $\beta_j = x_j(m'_j e)$. Since $m'_i e = M_{ii} e_i + \sum_{k \neq i} M_{ik} e_k$ and the errors are independent with mean zero,

$$\mathbb{E}[\alpha_i^2 | X] = M_{ii}^2 \mathbb{E}[e_i^4 | X] + \sigma_i^2 \sum_{k \neq i} M_{ik}^2 \sigma_k^2 \leq C, \quad \mathbb{E}[(m'_j e)^2 | X] = \sum_k M_{jk}^2 \sigma_k^2 \leq C M_{jj} \leq C,$$

where we used $\sum_k M_{ik}^2 = M_{ii} \leq 1$ and Assumption 2 (i). Hence $\mathbb{E}[\sum_i \bar{c}_i \alpha_i^2 | X] \leq C C_n$ and $\mathbb{E}[\sum_j \bar{c}_j \beta_j^2 | X] \leq C \max_j x_j^2 C_n$. By (i) the regressors have six moments, so $\max_j x_j^2 =$

$O_p(n^{1/3})$, and $C_n = O_p(1)$ was shown above; Markov's inequality therefore gives

$$\sum_{i=1}^n \bar{c}_i \alpha_i^2 = O_p(1), \quad \sum_{j=1}^n \bar{c}_j \beta_j^2 = O_p(n^{1/3}).$$

Consequently $\|\mathcal{G}_1\| = O_p(n^{1/6}) = o_p(n^{1/2})$. Using the same argument with $\alpha = \beta$ gives $\|\mathcal{G}_2\| = O_p(n^{1/3}) = o_p(n)$. Since $g_{2i} = 0$ the remaining two forms vanish, $\mathcal{G}_3 = \mathcal{G}_4 = 0$, so all four conditions in (9) hold.

Step 4. The alternative. We turn to Theorem 2. We first check the hypotheses of Lemma 3, which give the first two conditions of (12) and $B_n^{\text{cf}} = o(V)$, then the two estimation-error conditions, then the last condition of the theorem. Here $L = M$ is symmetric and idempotent, so $\|L\| \leq 1$; the sixth moments are Assumption 2(i); (8) was verified above; and $\xi_n = O_p(1)$ because Assumption 2(iv) bounds $\max_i |\pi_i|$ while $\max_i |\lambda_i| = \max_i |m'_i \omega| = o_p(1)$ by (iii).

For (A.1), recall $a_{ij}^2 = P_{ij}^2/K$ and $c_{ij} = (2/K)\tilde{P}_{ij}^2$ with $M_{ii}M_{jj} + M_{ij}^2 \geq c_M^2$. Since $\sum_{i=1}^n \sum_{j \neq i} P_{ij}^4 = o_p(K)$ and $A_n = 1 + o_p(1)$,

$$C_2 \leq \frac{C}{K^2} \sum_{i=1}^n \sum_{j \neq i} P_{ij}^4 = o_p(K^{-1}) = o_p(A_n^2),$$

and, because $\sum_i P_{ij}^2 = P_{jj} \leq \max_i P_{ii} = o_p(1)$,

$$\bar{c}_j \leq \frac{C}{K} \sum_i P_{ij}^2 = o_p(K^{-1}), \quad \text{so} \quad \max_j \bar{c}_j = o_p(A_n).$$

For the two Frobenius quantities, $M_{ij} = -P_{ij}$ for $i \neq j$ gives $L_F^2 = \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 \leq K$. Writing $M\Sigma M = \Sigma - P\Sigma - \Sigma P + P\Sigma P$, the off-diagonal entries of $M\Sigma M$ are $-(\sigma_i^2 + \sigma_j^2)P_{ij} + (P\Sigma P)_{ij}$, and

$$\|P\Sigma P\|_F^2 = \text{tr}(P\Sigma P P \Sigma P) \leq C^2 \text{tr}(P) = C^2 K,$$

so that $S_F \leq C\sqrt{\text{tr}(P)} = C\sqrt{K}$. Hence

$$C_2^{1/2}(L_F + S_F) = o_p(K^{-1/2})O_p(K^{1/2}) = o_p(1) = o_p(A_n),$$

and Lemma 3 delivers the first two conditions of (12) together with $B_n^{\text{cf}} = o(V)$.

It remains to bound the estimation error. For the cross-fit estimator, $m'_i X = 0$ by (ii), so Lemma 1 again gives $g_{2i} = 0$ and $\mathcal{R}_n = 0$, while now $g_{0i} = \check{A}_i$ and $g_{1i} = (m'_i \check{u})x_i$ with $m'_i \check{u} = m'_i e + \lambda_i$. Relative to the null the only changes are that $m'_i e$ acquires the shift λ_i and that $e_i(m'_i e)$ becomes $\check{A}_i$. Both factors keep the uniform second-moment bounds that (B.1) needs: $\mathbb{E}[\check{A}_i^2|X] \leq C(1 + \pi_i^2)(1 + \lambda_i^2) \leq C$ by Assumption 2(iv), and $\mathbb{E}[(m'_i \check{u})^2|X] \leq 2\lambda_i^2 + 2\mathbb{E}[(m'_i e)^2|X] \leq C$ because $\max_i |\lambda_i| = o_p(1)$. The computation used under the null therefore

applies unchanged and gives $\|\mathcal{G}_1\| = O_p(n^{1/6})$ and $\|\mathcal{G}_2\| = O_p(n^{1/3})$. With $\|\hat{\delta}\| = O_p(n^{-1/2})$ and $V \geq 2c^2 A_n$, this yields $\hat{V}_{\text{cf}} - \check{V}_{\text{cf}} = o_p(V)$.

For the plug-in estimator, (4) is (5) with ℓ_i the i th unit vector and $2a_{ij}^2$ in place of c_{ij} , so Lemma 1 applies with $L = I$ and gives $g_{0i} = \check{u}_i^2$, $g_{1i} = 2\check{u}_i x_i$ and $g_{2i} = x_i x_i'$; here the degree-three and degree-four terms do not vanish. The first two forms need different moment bounds, since g_{0i} and g_{1i} no longer carry $m'_i \check{u}$. Writing $\bar{a}_i = \sum_{j \neq i} 2a_{ij}^2$, so that $\sum_i \bar{a}_i = 2A_n = O_p(1)$, three sums bound them:

$$\sum_i \bar{a}_i \check{u}_i^4 = O_p(1), \quad \sum_i \bar{a}_i \check{u}_i^2 x_i^2 = O_p(n^{1/3}), \quad \sum_i \bar{a}_i x_i^4 = O_p(n^{2/3}),$$

the first two because $\mathbb{E}[\check{u}_i^4 | X] \leq C(1 + \pi_i^4) \leq C$ by (i) and Assumption 2(iv), the last two because $\max_i x_i^2 = O_p(n^{1/3})$. By (B.1) the first two give $\|\mathcal{G}_1\| = O_p(n^{1/6})$, while the first and third bound the $g_{0i}g_{2j}$ part of $\mathcal{G}_2$ and the second its $g_{1i}g'_{1j}$ part, so $\|\mathcal{G}_2\| = O_p(n^{1/3})$. Hence $\|\hat{\delta}\| \|\mathcal{G}_1\| + \|\hat{\delta}\|^2 \|\mathcal{G}_2\| = O_p(n^{-1/3})$. For the other two, $\max_i \|x_i\| = O_p(n^{1/6})$ and $\max_i |\check{u}_i| = O_p(b_n)$ by (i) and (iii), so $\mathcal{G}_3 = O_p(b_n n^{1/2})$ and $\mathcal{G}_4 = O_p(n^{2/3})$, whence $\|\hat{\delta}\|^3 \mathcal{G}_3 + \|\hat{\delta}\|^4 \mathcal{G}_4 = O_p(b_n n^{-1}) + O_p(n^{-4/3}) = o_p(1)$. Since $V + B_n \geq V \geq 2c^2 A_n$, this gives $\hat{V} - \check{V} = o_p(V + B_n)$.

Finally, the last condition of Theorem 2 is the non-degeneracy condition of Assumption 2(iv), so $\liminf_n B_n/V > 0$ and the plug-in estimator is inconsistent for V . All conditions of Theorems 1 and 2 therefore hold, which proves Proposition 1. $\square$

APPENDIX C. ASSUMPTIONS AND PROOF OF PROPOSITION 2

Assumption 3. *The following conditions hold, in the notation of Section 3.2.*

(i): *The errors $\{(u_i, v_i)\}_{i=1}^n$ are independent across i with $\mathbb{E}[u_i] = 0$ and $\mathbb{E}[v_i] = 0$, and satisfy $0 < c \leq \mathbb{E}[u_i^2] \leq C < \infty$, $\max_i \mathbb{E}|u_i|^6 \leq C$ and $\max_i \mathbb{E}\|v_i\|^6 \leq C$. The reduced form is bounded, $\max_i \|\Upsilon_i\| \leq C$. The error of the framework satisfies the same bounds: with $e_i = (y_i - \beta'_* x_i) - \bar{u}_i$,*

$$0 < c \leq \mathbb{E}[e_i^2] \leq C < \infty, \quad \max_i \mathbb{E}|e_i|^6 \leq C.$$

(ii): *The instruments Z are nonrandom with $\text{rank}(Z) = K$, the number of instruments satisfies $K \rightarrow \infty$ and $K/n \rightarrow \alpha \in [0, 1)$, and there is a constant c_P with $P_{ii} \leq c_P < 1$ for all i , and the reduced form is spanned by them, $\Upsilon = Z\Psi$.*

(iii): *The estimator is consistent for the pseudo-true value, $\hat{\beta} \xrightarrow{P} \beta_*$, and $\max_i |\omega_i| = o_p(b_n^{-3})$ for a deterministic sequence $1 \leq b_n = o(n)$ satisfying $\max_i |e_i| = O_p(b_n)$ and $\max_i |m'_i e| = O_p(b_n)$.*

(iv): Under the alternative the misspecification is bounded and non-degenerate, and the weights are not concentrated on a few pairs:

$$\max_i |\bar{u}_i| \leq C, \quad \liminf_n A_n^{-1} \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 \bar{u}_i^2 > 0, \quad \sum_{i=1}^n \sum_{j \neq i} P_{ij}^4 = o(K).$$

Under the null $\beta_* = \beta$ and $e_i = u_i$, so the bounds (i) places on e_i are vacuous; under the alternative $e_i = u_i + v'_i(\beta - \beta_*)$ and the bounds are a genuine restriction, since neither $\|\beta - \beta_*\|$ nor the lower bound on $\mathbb{E}[e_i^2]$ follows from the moments of u_i and v_i alone.

The first condition of (iv) is vacuous under the null, where $\bar{u} = 0$. The second makes the plug-in estimator inconsistent as required in Theorem 2. The third is implied by $\max_{i \neq j} |P_{ij}| = o(1)$.

Proof. For the overidentification-testing problem,

$$a_{ij} = P_{ij}, \quad c_{ij} = \frac{2P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2}, \quad \ell_i = m_i,$$

where $M = I - P$ and m_i denotes the i -th column of M . We verify the conditions of Theorems 1 and 2.

Step 1. The maintained assumption. We begin by verifying Assumption 1. By (ii), $P_{ii} \leq c_P < 1$ gives $M_{ii} = 1 - P_{ii} \geq 1 - c_P$, hence $M_{ii}M_{jj} + M_{ij}^2 \geq (1 - c_P)^2$ uniformly over $i \neq j$. Because P is idempotent,

$$\sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 = \text{tr}(P^2) - \sum_i P_{ii}^2 = K - \sum_i P_{ii}^2.$$

Moreover $\sum_i P_{ii}^2 \leq c_P \sum_i P_{ii} = c_P K$ by (ii), so that

$$(1 - c_P)K \leq \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 \leq K, \quad A_n = \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 = \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2.$$

Moreover,

$$2c^2 A_n \leq V = 2 \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \mathbb{E}[e_i^2] \mathbb{E}[e_j^2] \leq 2C^2 A_n,$$

because the variances are bounded above and away from zero. The leverage condition of Assumption 1 holds with $c_L = (1 - c_P)^2$, and

$$C_n = \sum_{i=1}^n \sum_{j \neq i} |c_{ij}| \leq \frac{2}{(1 - c_P)^2} A_n.$$

Thus $C_n = O(A_n)$. The maximal bounds for e_i , $m'_i e$, ω_i , and $m'_i \omega$ are imposed in Assumption 3(iii), the one for $m'_i \omega$ through $m'_i \omega = \omega_i$, which holds because $\omega = M\bar{u}$ and M is idempotent. This establishes Assumption 1.

Step 2. Orthogonality of the drift. We next verify the orthogonality condition appearing in Theorem 2. By the decomposition in Section 3.2, $\Delta = P\bar{u}$. Hence $\Delta \in \text{col}(Z)$. Since P is the projection matrix onto $\text{col}(Z)$ and $M = I - P$, $M\Delta = (I - P)\Delta = 0$. Therefore, $m'_i \Delta = 0$ for every i , and consequently $\max_i |m'_i \Delta| = 0 = o_p(b_n^{-3})$.

Step 3. The null hypothesis. First we check condition (8) for the error e_i of the framework, as Theorems 1 and 2 both require. Under H_0 the departure vanishes, $\bar{u} = 0$, so $\Delta = 0$, $\omega = 0$ and $e_i = u_i$: the null is the special case $e_i = u_i$ of what follows. Writing $\tilde{P}_{ij}^2 = P_{ij}^2 / (M_{ii}M_{jj} + M_{ij}^2)$, we have $c_{ij} = 2\tilde{P}_{ij}^2$ and $2a_{ij}^2 = 2P_{ij}^2$, so that

$$\sum_{i=1}^n \sum_{j \neq i} \{c_{ij} e_i(m'_i e) e_j(m'_j e) - 2a_{ij}^2 \mathbb{E}[e_i^2] \mathbb{E}[e_j^2]\} = 2K \cdot \frac{1}{K} \sum_{i=1}^n \sum_{j \neq i} \{\tilde{P}_{ij}^2 e_i(m'_i e) e_j(m'_j e) - P_{ij}^2 \mathbb{E}[e_i^2] \mathbb{E}[e_j^2]\}.$$

The sum on the right is the expression of Mikusheva and L. Sun (2022, Lemma 2) with P the projection onto $\text{col}(Z)$, $M = I - P$ and $U_i^{(1)} = U_i^{(2)} = U_i^{(3)} = e_i$. Their conditions hold by Assumption 3: $K \rightarrow \infty$ and $P_{ii} \leq c_P < 1$ by (ii), and by (i) the e_i are independent across i with mean zero and $\max_i \mathbb{E}|e_i|^6 \leq C$. This is where the bounds (i) places on e_i itself are used. Hence the right-hand side is $2K o_p(1) = o_p(K)$, which is $o_p(V)$ because $V \geq 2c^2 A_n \geq 2c^2(1 - c_P)K$. This is (8).

Next we check condition (9). Recall from (ii) that $\Upsilon = Z\Psi$, so $M\Upsilon = 0$ and $h_i = \ell'_i H = m'_i x = m'_i v$. We again give the argument for $G = 1$. Under H_0

$$g_{0i} = u_i(m'_i u), \quad g_{1i} = u_i(m'_i v) + (m'_i u)x_i, \quad g_{2i} = x_i(m'_i v).$$

Since $m'_i u = \sum_k M_{ik} u_k$ with $\sum_k M_{ik}^2 = M_{ii} \leq 1$, independence and the fourth moments of (i) give $\mathbb{E}[(m'_i u)^2] \leq C$ and $\mathbb{E}[(m'_i u)^4] \leq C$, and likewise for $m'_i v$. With $x_i = \Upsilon_i + v_i$ and $\max_i \|\Upsilon_i\| \leq C$ the Cauchy–Schwarz inequality then gives $\max_k \max_i \mathbb{E}\|g_{ki}\|^2 \leq C$.

With $\bar{c}_i = \sum_{j \neq i} |c_{ij}|$, so that $\sum_i \bar{c}_i = C_n$, and $c_{ij} = c_{ji}$ here, we can apply (B.1). Since $\mathbb{E}[\sum_{i=1}^n \bar{c}_i \|g_{ki}\|^2] \leq CC_n$, Markov's inequality gives $\sum_{i=1}^n \bar{c}_i \|g_{ki}\|^2 = O_p(C_n)$ for $k = 0, 1, 2$. By the same symmetry, the two terms of $\mathcal{G}_1$ coincide and so do the last two of $\mathcal{G}_2$, so that

$$\|\mathcal{G}_1\| \leq 2 \sum_{i=1}^n \sum_{j \neq i} |c_{ij}| |g_{0i}| \|g_{1j}\|, \quad \|\mathcal{G}_2\| \leq \sum_{i=1}^n \sum_{j \neq i} |c_{ij}| \|g_{1i}\| \|g_{1j}\| + 2 \sum_{i=1}^n \sum_{j \neq i} |c_{ij}| |g_{0i}| \|g_{2j}\|.$$

For $\mathcal{G}_3$ and $\mathcal{G}_4$, with $\|\delta\| = 1$ we have $|g'_{1i} \delta| \leq \|g_{1i}\|$ and $|\delta' g_{2i} \delta| \leq \|g_{2i}\|$, so the summand of $\mathcal{G}_3$ is at most $|c_{ij}|(\|g_{1i}\| \|g_{2j}\| + \|g_{1j}\| \|g_{2i}\|)$ and that of $\mathcal{G}_4$ at most $|c_{ij}| \|g_{2i}\| \|g_{2j}\|$. Summing

these and using the symmetry of c once more,

$$\mathcal{G}_3 \leq 2 \sum_{i=1}^n \sum_{j \neq i} |c_{ij}| \|g_{1i}\| \|g_{2j}\|, \quad \mathcal{G}_4 \leq \sum_{i=1}^n \sum_{j \neq i} |c_{ij}| \|g_{2i}\| \|g_{2j}\|.$$

By (B.1) each of the five double sums on the right is $O_p(C_n)$. Since $C_n \leq 2A_n/(1 - c_P)^2 \leq 2K/(1 - c_P)^2$, all $\mathcal{G}_k$ are $O_p(K)$ for $k = 1, 2, 3, 4$.

Consistency of $\hat{\beta}$ gives $\|\hat{\delta}\| = o_p(1)$, and $V \geq 2c^2 A_n \geq 2c^2(1 - c_P)K$, so

$$\|\hat{\delta}\| \|\mathcal{G}_1\| + \|\hat{\delta}\|^2 \|\mathcal{G}_2\| + \|\hat{\delta}\|^3 \mathcal{G}_3 + \|\hat{\delta}\|^4 \mathcal{G}_4 = o_p(K) = o_p(V),$$

which is (9). All conditions of Theorem 1 are therefore satisfied, which proves the null part of Proposition 2.

Step 4. The alternative. We turn to Theorem 2. We first check the hypotheses of Lemma 3, which give the first two conditions of (12) and $B_n^{\text{cf}} = o(V)$, then the two estimation-error conditions, then the last condition of the theorem. Because $\Delta = P\bar{u}$ and $\omega = M\bar{u}$, the systematic component is the misspecification itself and the surviving part is its projection off the instruments,

$$\pi_i = \Delta_i + \omega_i = \bar{u}_i, \quad \lambda_i = m'_i \omega = (M\bar{u})_i = \omega_i,$$

the last equality because M is idempotent.

We check the hypotheses of Lemma 3. Again $L = M$ is symmetric idempotent, so $\|L\| \leq 1$; the sixth moments are Assumption 3(i); (8) was verified above; and $\xi_n = O_p(1)$ by Assumption 3(iv) together with $\max_i |\lambda_i| = \max_i |\omega_i| = o_p(1)$ from (iii). For (A.1), $c_{ij} = 2\tilde{P}_{ij}^2$ with $M_{ii}M_{jj} + M_{ij}^2 \geq (1 - c_P)^2$, so $\mathcal{C}_2 \leq C \sum_{i=1}^n \sum_{j \neq i} P_{ij}^4 = o_p(K)$ by Assumption 3(iv), which is $o_p(A_n^2)$ because $A_n \geq (1 - c_P)K \rightarrow \infty$. Next $\bar{c}_j \leq C \sum_i P_{ij}^2 = CP_{jj} \leq Cc_P$, so $\max_j \bar{c}_j = O(1) = o(A_n)$. Finally $L_F^2 = \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 \leq K$ and, as in the proof of Proposition 1, $S_F \leq C\sqrt{K}$; hence

$$\mathcal{C}_2^{1/2}(L_F + S_F) = o_p(K^{1/2})O(K^{1/2}) = o_p(K) = o_p(A_n).$$

Lemma 3 therefore gives the first two conditions of (12) and $B_n^{\text{cf}} = o(V)$. Note where the conditions of Assumption 3(iv) act: the bound on $\bar{u}$ controls the size of the systematic component, and the summability of P_{ij}^4 stops a single pair of observations from carrying a non-negligible share of the weight.

The estimation-error conditions are those of the null argument with u_i replaced by $\check{u}_i = \bar{u}_i + e_i$ and $m'_i u$ by $m'_i \check{u} = \lambda_i + m'_i e$. Since $\xi_n = O_p(1)$, the moment bounds $\max_k \max_i \mathbb{E} \|g_{ki}\|^2 \leq C$

used there continue to hold, so

$$\|\hat{\delta}\|\|\mathcal{G}_1\| + \|\hat{\delta}\|^2\|\mathcal{G}_2\| + \|\hat{\delta}\|^3\mathcal{G}_3 + \|\hat{\delta}\|^4\mathcal{G}_4 = o_p(K) = o_p(V);$$

this gives $\hat{V}_{\text{cf}} - \check{V}_{\text{cf}} = o_p(V)$. For the plug-in estimator, (4) is (5) with ℓ_i the i th unit vector and $2a_{ij}^2 = 2P_{ij}^2$ in place of c_{ij} , so Lemma 1 applies with $L = I$ and gives $g_{0i} = \check{u}_i^2$, $g_{1i} = 2\check{u}_i x_i$, $g_{2i} = x_i x_i'$, the weights summing to $2A_n \leq 2K$. The same moment bound $\max_k \max_i \mathbb{E}\|g_{ki}\|^2 \leq C$ holds for these, using $\max_i \|\Upsilon_i\| \leq C$ and the sixth moments of v_i , so

$$\|\hat{\delta}\|\|\mathcal{G}_1\| + \|\hat{\delta}\|^2\|\mathcal{G}_2\| + \|\hat{\delta}\|^3\mathcal{G}_3 + \|\hat{\delta}\|^4\mathcal{G}_4 = o_p(K) = o_p(V),$$

and hence $\hat{V} - \check{V} = o_p(V) = o_p(V + B_n)$.

Finally, Assumption 3(iv) requires $\liminf_n A_n^{-1} \sum_{i=1}^n \sum_{j \neq i} P_{ij}^2 \bar{u}_i^2 > 0$, which makes the plug-in estimator inconsistent as required in Theorem 2. All conditions of Theorems 1 and 2 therefore hold, which proves Proposition 2. $\square$

APPENDIX D. ASSUMPTIONS AND PROOF OF PROPOSITION 3

Assumption 4. *The following conditions hold, in the notation of Section 3.3.*

(i): $\{(y_i, w_i, z_i)\}_{i=1}^n$ are i.i.d. and

$$y_i = w_i' \beta_1 + z_i' \beta_2 + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | x_i] = 0,$$

where $x_i = (w_i', z_i')'$. Also,

$$0 < c \leq \sigma_i^2 := \mathbb{E}[\varepsilon_i^2 | x_i] \leq C < \infty, \quad \max_{1 \leq i \leq n} \mathbb{E}[\varepsilon_i^6 | x_i] \leq C,$$

uniformly in i , for some constants $0 < c < C < \infty$. In addition, $\mathbb{E}\|w_i\|^2 < \infty$ and $\mathbb{E}[w_i w_i']$ is nonsingular.

(ii): The number of restrictions satisfies $r \rightarrow \infty$, and there is a constant $c_M > 0$ such that $\min_i M_{ii} \geq c_M$ with probability approaching one.

(iii): There exists a deterministic sequence $1 \leq b_n = o(n)$ such that $\max_i |\varepsilon_i| = O_p(b_n)$ and $\max_i |m_i' \varepsilon| = O_p(b_n)$, where m_i denotes the i -th column of M .

(iv): Under the alternative the drift is bounded and non-degenerate, and the weights are not concentrated on a few pairs:

$$\max_i |\Delta_i| = O_p(1), \quad \liminf_n A_n^{-1} \sum_{i=1}^n \sum_{j \neq i} Q_{ij}^2 \Delta_i^2 > 0, \quad \sum_{i=1}^n \sum_{j \neq i} Q_{ij}^4 = o_p(r).$$

The condition of Anatolyev and Solvsten (2023) is the fourth-moment bound $\max_i \mathbb{E}[\varepsilon_i^4 | x_i] + \sigma_i^{-2} = O_p(1)$; Mikusheva and L. Sun (2022, Lemma 2) requires the sixth moment, and we

state the bounds with deterministic constants so that they supply Assumption 1 directly, as in Assumptions 2 and 3 above. For the same reason (ii) fixes the constant c_M . Their $\max_i M_{ii}^{-1} = O_p(1)$ gives a bound that may depend on how small a failure probability one insists on, whereas here c_M has to be a single constant: it becomes the $c_L = c_M^2$ of Assumption 1, and the proof carries it on through $c_P = 1 - c_M$ and $A_n \geq (1 - c_P)r$. The two conditions on w_i are not part of their Assumption 2.

We place no direct restriction on r/n : as the proof below shows, the condition in (ii) already delivers $\limsup_n r/n < 1$, the leverage bound $(P_X)_{ii} \leq c_P < 1$ that Mikusheva and L. Sun (2022, Lemma 2) requires of the projection to which it is applied, together with the bound $Q_{ii} \leq c_P$ used for $\sum_i Q_{ii}^2 \leq c_P r$, and two-sided bounds on A_n . Anatolyev and Sølvesten (2023) allow r to remain fixed; we retain $r \rightarrow \infty$, which enters through Mikusheva and L. Sun (2022, Assumption 1) and lets us avoid the non-standard limit that a fixed number of restrictions produces.

The first condition of (iv) is vacuous under the null, where $\Delta = 0$. The second makes the plug-in estimator inconsistent as required in Theorem 2. Note that the first condition is stated for $\Delta = M_W Z(\beta_2 - \beta_{20})$ and does not follow from a bound on $z'_i(\beta_2 - \beta_{20})$ alone, since $(M_W \Delta)_i = \Delta_{0i} - w'_i(W'W)^{-1}W'\Delta_0$ and (i) bounds $\max_i \|w_i\|$ only by $o_p(n^{1/2})$. The third is implied by $\max_{i \neq j} |Q_{ij}| = o_p(1)$.

Proof. For the many restrictions testing problem,

$$a_{ij} = Q_{ij}, \quad c_{ij} = 2 \frac{Q_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2}, \quad \ell_i = m_i,$$

where m_i denotes the i -th column of M . We verify the conditions of Theorems 1 and 2. All expectations below are conditional on $X = [W, Z]$.

Step 1. The maintained assumption. We begin by verifying Assumption 1. By Assumption 4(ii), $\min_i M_{ii} \geq c_M > 0$ with probability approaching one, so that $M_{ii}M_{jj} + M_{ij}^2 \geq c_M^2$ uniformly over $i \neq j$, the leverage condition of Assumption 1 with $c_L = c_M^2$. The same bound on $\min_i M_{ii}$ carries the remaining leverage and order calculations of this step. It first bounds the leverages of Q , which supplies the weights $a_{ij} = Q_{ij}$: by the Frisch–Waugh–Lovell decomposition, $P_X = P_W + Q$ with $P_W = W(W'W)^{-1}W'$ positive semidefinite, so that $(P_W)_{ii} \geq 0$ and

$$Q_{ii} \leq Q_{ii} + (P_W)_{ii} = (P_X)_{ii} = 1 - M_{ii} \leq 1 - c_M =: c_P < 1.$$

Averaging $1 - M_{ii} \leq 1 - c_M$ over i and using $\sum_i (1 - M_{ii}) = d$ gives $d/n \leq 1 - c_M$; since $r \leq d$ this makes $\limsup_n r/n \leq 1 - c_M < 1$. Finally Q is symmetric idempotent of rank r ,

so $\sum_{i=1}^n \sum_{j \neq i} Q_{ij}^2 = \text{tr}(Q^2) - \sum_i Q_{ii}^2 = r - \sum_i Q_{ii}^2$. Moreover $\sum_i Q_{ii}^2 \leq c_P \sum_i Q_{ii} = c_P r$, so that

$$(1 - c_P)r \leq A_n = \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 = \sum_{i=1}^n \sum_{j \neq i} Q_{ij}^2 \leq r.$$

Since $0 < c \leq \sigma_i^2 \leq C < \infty$, we obtain

$$2c^2 A_n \leq V = 2 \sum_{i=1}^n \sum_{j \neq i} a_{ij}^2 \sigma_i^2 \sigma_j^2 \leq 2C^2 A_n.$$

Furthermore,

$$C_n = \sum_{i=1}^n \sum_{j \neq i} |c_{ij}| \leq \frac{2}{c_M^2} \sum_{i=1}^n \sum_{j \neq i} Q_{ij}^2 = \frac{2}{c_M^2} A_n.$$

Assumption 4(iii) supplies the maximal bounds for ε_i and $m'_i \varepsilon$. The approximation error vanishes here, $\omega = 0$, so the conditions Assumption 1 places on ω hold trivially. Write $\hat{\delta} = -(n^{-1}W'W)^{-1}(n^{-1}W'\varepsilon)$ and bound the two factors. By (i) the matrix $\mathbb{E}[w_i w_i']$ is nonsingular, so $\lambda_{\min}(n^{-1}W'W)$ is bounded away from zero with probability approaching one; and since $\mathbb{E}[\varepsilon_i | x_i] = 0$ kills the cross terms, $\mathbb{E}\|n^{-1}W'\varepsilon\|^2 = n^{-2} \sum_i \mathbb{E}\|w_i \varepsilon_i\|^2 \leq C n^{-1} \mathbb{E}\|w_i\|^2$, which is $O(n^{-1})$ by (i). Hence $\|\hat{\delta}\| = O_p(n^{-1/2})$ under both hypotheses. Moreover $m'_i \rho = m'_i W \hat{\delta} = 0$ identically because $MW = 0$. This establishes Assumption 1.

Step 2. Orthogonality of the drift. We next verify the orthogonality condition in Theorem 2. Under the alternative, $\Delta = M_W Z(\beta_2 - \beta_{20})$. Since $Z(\beta_2 - \beta_{20}) \in \text{col}(Z) \subseteq \text{col}(X)$ and $MW = 0$ gives $MP_W = 0$, it follows that $M\Delta = 0$. Therefore $m'_i \Delta = 0$ for every i , and hence $\max_i |m'_i \Delta| = 0 = o_p(b_n^{-3})$.

Step 3. The null hypothesis. First we check condition (8). Under H_0 we have $\beta_2 = \beta_{20}$, so $\Delta = 0$, and $\omega = 0$, so $\tilde{u}_i = \varepsilon_i$ in the notation of Lemma 1. With $\tilde{P}_{ij}^2 = Q_{ij}^2 / (M_{ii}M_{jj} + M_{ij}^2)$ we have $c_{ij} = 2\tilde{P}_{ij}^2$ and $2a_{ij}^2 = 2Q_{ij}^2$, so

$$\sum_{i=1}^n \sum_{j \neq i} \{c_{ij} \varepsilon_i(m'_i \varepsilon) \varepsilon_j(m'_j \varepsilon) - 2a_{ij}^2 \sigma_i^2 \sigma_j^2\} = 2 \sum_{i=1}^n \sum_{j \neq i} \{\tilde{P}_{ij}^2 \varepsilon_i(m'_i \varepsilon) \varepsilon_j(m'_j \varepsilon) - Q_{ij}^2 \sigma_i^2 \sigma_j^2\}. \quad (\text{D.1})$$

Mikusheva and L. Sun (2022, Lemma 2) is stated for a projection and its own complement, and Q is not the complement of M : here $M = I - P_X$ with $P_X = P_W + Q$, so the lemma does not apply to the pair (Q, M) directly. It does apply to the pair (P_X, M) , and we reduce to that case by splitting the weight. Rearranging the decomposition of Step 1 as $Q = P_X - P_W$ gives

$$Q_{ij}^2 = (P_X)_{ij}^2 - 2(P_X)_{ij}(P_W)_{ij} + (P_W)_{ij}^2.$$

Substituting this into the right-hand side of (D.1) yields three pieces

$$2 \sum_{i=1}^n \sum_{j \neq i} \left\{ \frac{v_{ij}}{M_{ii}M_{jj} + M_{ij}^2} \varepsilon_i(m'_i \varepsilon) \varepsilon_j(m'_j \varepsilon) - v_{ij} \sigma_i^2 \sigma_j^2 \right\}, \quad \text{for } v_{ij} \in \{(P_X)_{ij}^2, -2(P_X)_{ij}(P_W)_{ij}, (P_W)_{ij}^2\}.$$

For $v_{ij} = (P_X)_{ij}^2$, it is $2d$ times the expression of Mikusheva and L. Sun (2022, Lemma 2) with $P = P_X$ and $U_i^{(1)} = U_i^{(2)} = U_i^{(3)} = \varepsilon_i$. Their conditions hold: P_X is a projection of rank $d \rightarrow \infty$, its complement is M , the leverage bound $(P_X)_{ii} = 1 - M_{ii} \leq 1 - c_M < 1$ holds by (ii), and the errors are independent with mean zero and six bounded conditional moments by (i). Hence it is $2d o_p(1) = o_p(d) = o_p(r)$.

For $v_{ij} = -2(P_X)_{ij}(P_W)_{ij}$ and $(P_W)_{ij}^2$, the pieces are negligible because w_i has fixed dimension, so that $\text{tr}(P_W) = d - r$ stays bounded. For a symmetric idempotent matrix A one has $\sum_i \sum_j A_{ij}^2 = \text{tr}(A)$, so by the Cauchy–Schwarz inequality

$$\begin{aligned} \sum_{i=1}^n \sum_{j \neq i} |(P_X)_{ij}(P_W)_{ij}| &\leq \{\text{tr}(P_X) \text{tr}(P_W)\}^{1/2} = \{d(d-r)\}^{1/2} = O(r^{1/2}), \\ \sum_{i=1}^n \sum_{j \neq i} (P_W)_{ij}^2 &\leq d - r = O(1). \end{aligned}$$

Moreover, since $\sum_k M_{ik}^2 = M_{ii} \leq 1$, it follows

$$\mathbb{E}[\varepsilon_i^2(m'_i \varepsilon)^2 | X] = M_{ii}^2 \mathbb{E}[\varepsilon_i^4 | x_i] + \sigma_i^2 \sum_{k \neq i} M_{ik}^2 \sigma_k^2 \leq C.$$

Thus applying the Cauchy–Schwarz inequality gives $\mathbb{E}[|\varepsilon_i(m'_i \varepsilon) \varepsilon_j(m'_j \varepsilon)| | X] \leq C$ uniformly in $i \neq j$. Hence, for each of the two,

$$\mathbb{E} \left[\left| 2 \sum_{i=1}^n \sum_{j \neq i} \left\{ \frac{v_{ij}}{M_{ii}M_{jj} + M_{ij}^2} \varepsilon_i(m'_i \varepsilon) \varepsilon_j(m'_j \varepsilon) - v_{ij} \sigma_i^2 \sigma_j^2 \right\} \right| \middle| X \right] \leq C \sum_{i=1}^n \sum_{j \neq i} |v_{ij}|,$$

the first term inside the sum by $M_{ii}M_{jj} + M_{ij}^2 \geq c_M^2$ and the moment bound just established, the second by $\sigma_i^2 \leq C$. The right-hand side is $O(r^{1/2})$ for $v_{ij} = -2(P_X)_{ij}(P_W)_{ij}$ and $O(1)$ for $v_{ij} = (P_W)_{ij}^2$, and it is deterministic, so both pieces are $O_p(r^{1/2}) = o_p(r)$ by Markov's inequality. Hence the left-hand side is $o_p(r)$, which is $o_p(V)$ because $V \geq 2c^2 A_n \geq 2c^2(1 - c_P)r$. This is (8).

Next we check condition (9). The annihilation $\ell'_i H = m'_i W = 0$ gives $g_{2i} = 0$, so $\mathcal{G}_3 = \mathcal{G}_4 = \mathcal{R}_n = 0$ and only $\mathcal{G}_1$ and $\mathcal{G}_2$ remain. Under H_0 , $g_{0i} = \varepsilon_i(m'_i \varepsilon)$ and $g_{1i} = (m'_i \varepsilon)w_i$, exactly as in the proof of Proposition 1 with w_i in place of x_i . The same bound (B.1) applies: with

$\bar{c}_i = \sum_{j \neq i} |c_{ij}|$ and $\alpha_i = \varepsilon_i(m'_i \varepsilon)$, $\beta_j = w_j(m'_j \varepsilon)$,

$$\mathbb{E} \left[\sum_{i=1}^n \bar{c}_i \alpha_i^2 \mid X \right] \leq C C_n, \quad \mathbb{E} \left[\sum_{j=1}^n \bar{c}_j \|\beta_j\|^2 \mid X \right] \leq C \max_j \|w_j\|^2 C_n,$$

so that $\|\mathcal{G}_1\|$ and $\|\mathcal{G}_2\|$ are $O_p(C_n \max_j \|w_j\|)$ and $O_p(C_n \max_j \|w_j\|^2)$ respectively. Since w_i is fixed-dimensional with bounded second moments, $\max_j \|w_j\| = o_p(n^{1/2})$, and $C_n = O_p(A_n) = O_p(r)$, so with $\|\hat{\delta}\| = O_p(n^{-1/2})$

$$\|\hat{\delta}\| \|\mathcal{G}_1\| = o_p(r), \quad \|\hat{\delta}\|^2 \|\mathcal{G}_2\| = o_p(r),$$

both of which are $o_p(V)$. This is (9). All conditions of Theorem 1 are therefore satisfied, which proves the null part of Proposition 3.

Step 4. The alternative. We turn to Theorem 2. We first check the hypotheses of Lemma 3, which give the first two conditions of (12) and $B_n^{\text{cf}} = o(V)$, then the two estimation-error conditions, then the last condition of the theorem. Here $\omega = 0$, so the systematic component is the drift itself and the surviving part vanishes identically,

$$\pi_i = \Delta_i = (M_W Z(\beta_2 - \beta_{20}))_i, \quad \lambda_i = m'_i \omega = 0.$$

By Step 2, $M\Delta = 0$. This is the application in which the cross-fit bias reduces furthest: by Lemma 2 only the terms carrying $(M\Sigma M)_{ij}$ and M_{ij} survive.

We check the hypotheses of Lemma 3. As before $L = M$ is symmetric idempotent, so $\|L\| \leq 1$; the sixth moments are Assumption 4(i); (8) was verified above; and $\xi_n = \max_i |\Delta_i| = O_p(1)$ by Assumption 4(iv), λ being zero. For (A.1), $c_{ij} = 2Q_{ij}^2 / (M_{ii}M_{jj} + M_{ij}^2)$ with $M_{ii}M_{jj} + M_{ij}^2 \geq c_M^2$, so Assumption 4(iv) gives $\mathcal{C}_2 \leq C \sum_{i=1}^n \sum_{j \neq i} Q_{ij}^4 = o_p(r)$, which is $o_p(A_n^2)$ because $A_n \geq (1 - c_P)r \rightarrow \infty$. Next $\bar{c}_j \leq C \sum_i Q_{ij}^2 = CQ_{jj} \leq Cc_P$, so $\max_j \bar{c}_j = O(1) = o(A_n)$. For the Frobenius quantities, $\text{tr}(P_X) = d$ and $M_{ij} = -(P_X)_{ij}$ for $i \neq j$; then $L_F^2 \leq d$ and, exactly as in the proof of Proposition 1 with P_X in place of P , $S_F \leq C\sqrt{d}$. Since $d = r + \dim(w_i)$ and $r \rightarrow \infty$, $d/r \rightarrow 1$, whence

$$\mathcal{C}_2^{1/2}(L_F + S_F) = o_p(r^{1/2})O(d^{1/2}) = o_p(r) = o_p(A_n).$$

Lemma 3 therefore gives the first two conditions of (12) and $B_n^{\text{cf}} = o(V)$.

The estimation error is the easiest of the three applications, because the rate $\|\hat{\delta}\| = O_p(n^{-1/2})$ established above holds under the alternative as well as under the null. Since $MW = 0$ we have $\ell'_i H = m'_i W = 0$, so Lemma 1 gives $g_{2i} = 0$ and $\mathcal{R}_n = 0$, while $g_{0i} = \check{A}_i$ and $g_{1i} = (m'_i \varepsilon)w_i$; indeed $m'_i \hat{u} = m'_i \varepsilon$ exactly, under the alternative as under the null, because $M\Delta = 0$ and $MW = 0$. Write $\bar{c}_i = \sum_{j \neq i} |c_{ij}| \leq C \sum_j Q_{ij}^2 = CQ_{ii} \leq Cc_P$, so that

$\sum_i \bar{c}_i = C_n = O_p(r)$. By (B.1),

$$\|\mathcal{G}_1\| \leq C \left(\sum_i \bar{c}_i \check{A}_i^2 \right)^{1/2} \left(\sum_j \bar{c}_j (m'_j \varepsilon)^2 \|w_j\|^2 \right)^{1/2}, \quad \|\mathcal{G}_2\| \leq C \sum_j \bar{c}_j (m'_j \varepsilon)^2 \|w_j\|^2.$$

Now $\mathbb{E}[\check{A}_i^2|X] \leq C(1+\Delta_i^2) \leq C$ by Assumption 4(iv) and $\mathbb{E}[(m'_i \varepsilon)^2|X] \leq C$ because $\sum_k M_{ik}^2 = M_{ii} \leq 1$, so the first sum is $O_p(C_n) = O_p(r)$ and the second is at most $C \max_j \bar{c}_j \sum_j \|w_j\|^2 = O_p(n)$, using $\mathbb{E}\|w_i\|^2 < \infty$. Hence $\|\hat{\delta}\| \|\mathcal{G}_1\| = O_p(n^{-1/2}) O_p(\sqrt{rn}) = O_p(\sqrt{r})$ and $\|\hat{\delta}\|^2 \|\mathcal{G}_2\| = O_p(1)$; since $V \geq 2c^2 A_n \geq 2c^2(1 - c_P)r \rightarrow \infty$, both are $o_p(V)$ and $\hat{V}_{\text{cf}} - \check{V}_{\text{cf}} = o_p(V)$.

For the plug-in estimator, (4) is (5) with ℓ_i the i th unit vector and $2a_{ij}^2 = 2Q_{ij}^2$ in place of c_{ij} , so Lemma 1 applies with $L = I$ and gives $g_{0i} = \check{u}_i^2$, $g_{1i} = 2\check{u}_i w_i$ and $g_{2i} = w_i w_i'$, the weights having row sums $\bar{a}_i = 2(Q_{ii} - Q_{ii}^2) \leq 2c_P$ and $\sum_i \bar{a}_i = 2A_n \leq 2r$. Three sums bound all four forms:

$$\sum_i \bar{a}_i \check{u}_i^4 = O_p(r), \quad \sum_i \bar{a}_i \check{u}_i^2 \|w_i\|^2 = O_p(n), \quad \sum_i \bar{a}_i \|w_i\|^4 = o_p(n^2),$$

the first because $\mathbb{E}[\check{u}_i^4|X] \leq C$ by Assumption 4(i) and (iv) while $\sum_i \bar{a}_i \leq 2r$, the second because $\bar{a}_i \leq 2c_P$ and $\mathbb{E}\|w_i\|^2 < \infty$, and the third because $\sum_i \bar{a}_i \|w_i\|^4 \leq 2c_P \max_i \|w_i\|^2 \sum_i \|w_i\|^2$ with $\max_i \|w_i\|^2 = o_p(n)$. The first two bound $\mathcal{G}_1$; the second bounds the $g_{1i} g'_{1j}$ part of $\mathcal{G}_2$ and the first and third its $g_{0i} g_{2j}$ part; the second and third bound $\mathcal{G}_3$, and the third bounds $\mathcal{G}_4$. Applying (B.1) to $\mathcal{G}_1, \dots, \mathcal{G}_4$ gives $\|\mathcal{G}_1\| = O_p(\sqrt{rn})$, $\|\mathcal{G}_2\| = O_p(n) + o_p(n\sqrt{r})$, $\mathcal{G}_3 = o_p(n^{3/2})$ and $\mathcal{G}_4 = o_p(n^2)$, whence

$$\|\hat{\delta}\| \|\mathcal{G}_1\| + \|\hat{\delta}\|^2 \|\mathcal{G}_2\| + \|\hat{\delta}\|^3 \mathcal{G}_3 + \|\hat{\delta}\|^4 \mathcal{G}_4 = O_p(\sqrt{r}) + o_p(\sqrt{r}) + o_p(1) = o_p(r),$$

so $\hat{V} - \check{V} = o_p(V) = o_p(V + B_n)$.

Finally the last condition of Theorem 2 is Assumption 4(iv). All conditions of Theorems 1 and 2 therefore hold, which proves Proposition 3. $\square$

APPENDIX E. SIZE-CORRECTED POWER FOR SECTION 5.3

We report size-corrected power for Section 5.3 because the raw sizes at $\eta = 0$ differ slightly across procedures, all between 3.6% and 4.4%. Size correction removes any advantage the cross-fit procedure enjoys merely from being better sized, and isolates the drift-contamination mechanism. Throughout, rejection frequencies use only the replications with a non-negative variance estimate.

Figure E.1 shows the size-corrected curves at the baseline sample size. The cross-fit advantage survives the correction: for $\phi(X) = \text{age}^3$ the peak cross-fit – plug-in gap is +5.0pp at $\eta = 0.15$, and cross-fit – AS23 is +4.8pp there; for the heavier-tailed $\phi(X) = \text{age}^5$ the two

gaps reach +6.6pp and +4.8pp. That the gap persists after equalizing size shows it reflects the reduction in drift contamination rather than the mild conservativeness of the plug-in. By $\eta = 0.30$ all three procedures approach saturation and the gaps close.

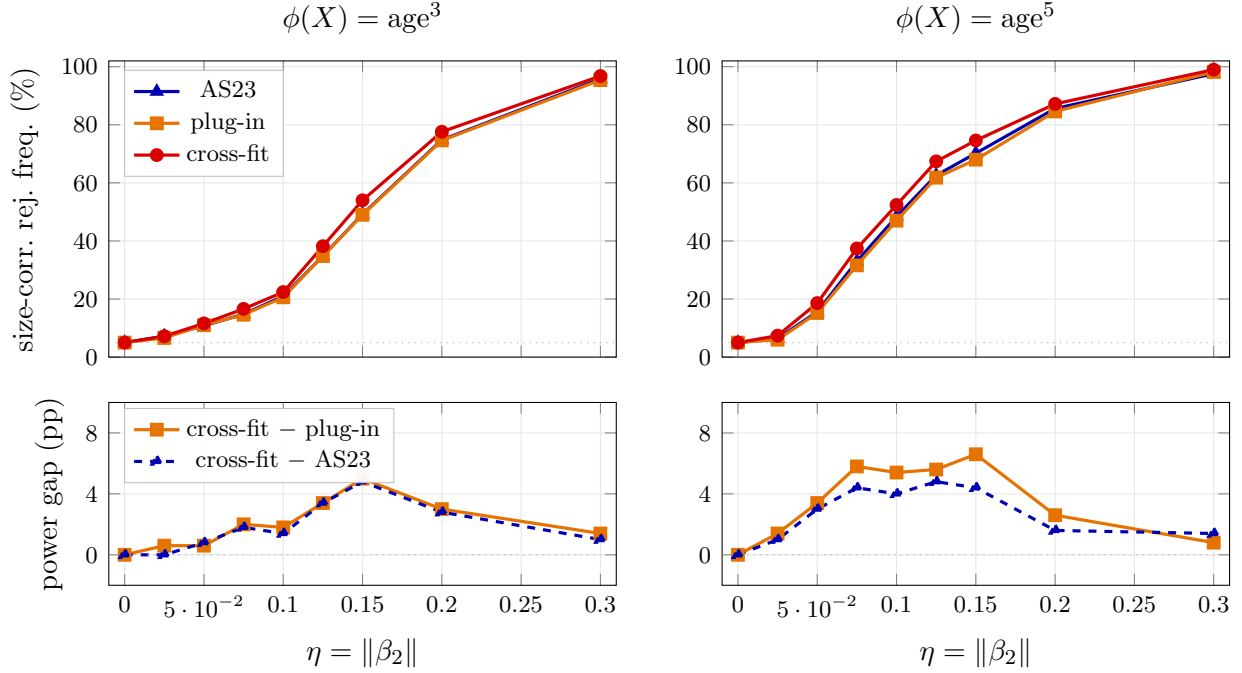


FIGURE E.1. Size-corrected power for the three many-restriction procedures (AS23, plug-in, cross-fit) under two heavy-tail HTE directions: $\phi(X) = \text{age}^3$ (left, $r = 27$) and $\phi(X) = \text{age}^5$ (right, $r = 28$). Bernoulli-LPM errors, $n = 1500$, 500 repetitions per η . Top row: size-corrected rejection frequency vs. $\eta = \|\beta_2\|$. Bottom row: the power gap of cross-fit over plug-in and over AS23. Rejection frequencies use only the repetitions with non-negative variance.

Table E.1 and Figure E.2 repeat the exercise across the four sample sizes. The table reports frequencies at the representative alternative $\eta = 0.15$, together with the largest number of negative variance estimates observed at any point of the η -grid; the figure gives the differences over the whole grid. The $n = 1,500$ column of the table is the baseline gap above. Rejection frequencies fall as n decreases at the fixed signal scale $\eta = 0.15$, since the aggregate signal in the statistic weakens. The gap over the plug-in procedure depends on the shape of the alternative: under the cubic alternative it falls from 5.0 percentage points at $r/n = 1.8\%$ to about zero at 5.4%, while under the fifth-order alternative it stays between 4.0 and 7.0 points.

TABLE E.1. Size-corrected rejection frequencies as r/n increases

Design	r/n	Rejection frequency at $\eta = 0.15$			Difference		Max. negative estimates	
		AS23	Plug-in	Cross-fit	CF-plug-in	CF-AS23	AS23	Cross-fit
age^3 , $n = 1,500$	1.8%	0.492	0.490	0.540	0.050	0.048	19	0
age^3 , $n = 1,000$	2.7%	0.270	0.256	0.294	0.038	0.024	17	0
age^3 , $n = 750$	3.6%	0.207	0.220	0.236	0.016	0.029	18	0
age^3 , $n = 500$	5.4%	0.136	0.160	0.162	0.002	0.026	26	0
age^5 , $n = 1,500$	1.8%	0.702	0.680	0.746	0.066	0.044	26	0
age^5 , $n = 1,000$	2.8%	0.510	0.496	0.566	0.070	0.056	33	0
age^5 , $n = 750$	3.7%	0.323	0.334	0.382	0.048	0.059	40	0
age^5 , $n = 500$	5.6%	0.189	0.178	0.218	0.040	0.029	50	0

Notes: Rejection frequencies are size-corrected and evaluated at $\eta = 0.15$. The final two columns report the maximum, over the reported η -grid, number of negative variance estimates among 500 replications.

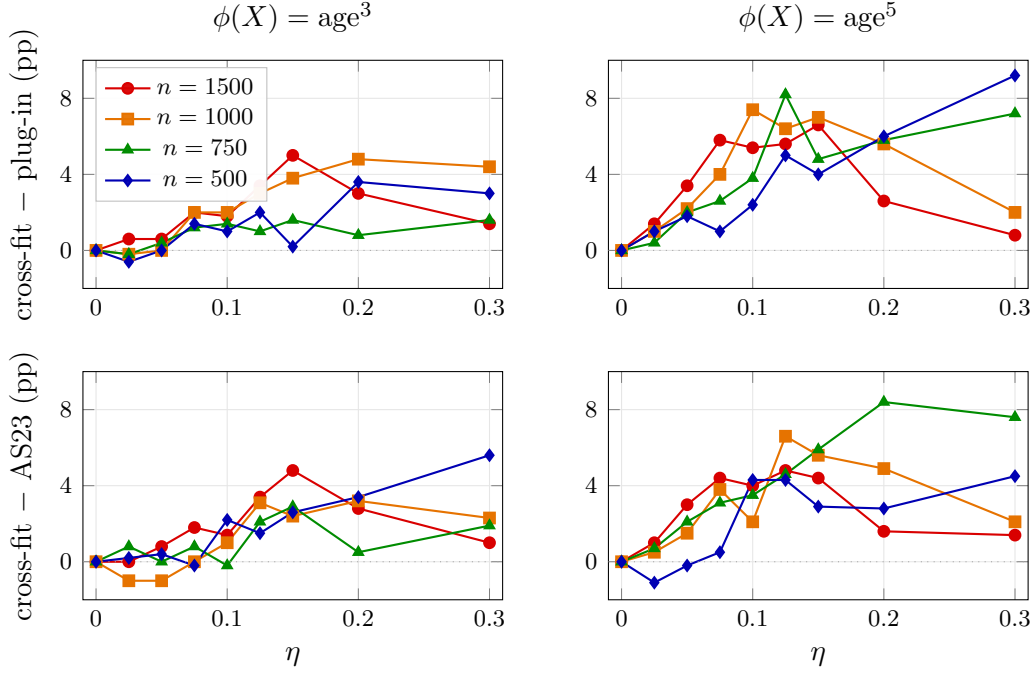


FIGURE E.2. Size-corrected rejection-frequency differences as r/n increases. The columns correspond to $\phi(X) = \text{age}^3$ ($r = 27$) and $\phi(X) = \text{age}^5$ ($r = 28$). The top row reports the cross-fit rejection frequency minus the plug-in rejection frequency, and the bottom row reports the cross-fit rejection frequency minus the Anatolyev–Sølvsten rejection frequency. Results are based on 500 Monte Carlo replications for each pair (η, n) and are computed over replications in which the corresponding variance estimates are nonnegative.

ACKNOWLEDGEMENTS

We are grateful to Mikkel Sølvsten and Luke Taylor for helpful comments. Matsushita gratefully acknowledges financial support from JSPS KAKENHI (23K01331); Sunada gratefully acknowledges research support from the Aarhus Center for Econometrics (ACE), funded by the Danish National Research Foundation grant number DNRF186.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE MANUSCRIPT PREPARATION PROCESS

During the preparation of this work the authors used Claude Opus in order to assist with coding, formatting of tables and outputs, spelling, and proof-reading. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

- Anatolyev, S. and N. Gospodinov (2011). “Specification testing in models with many instruments”. *Econometric Theory* 27.2, 427–441. DOI: [10.1017/S0266466610000307](https://doi.org/10.1017/S0266466610000307).
- Anatolyev, S. and A. Korobka (2026). “Parameter-invariant unbiased estimation of individual variances and their pairwise products”. *Economics Letters* 268, 113145. DOI: [10.1016/j.econlet.2026.113145](https://doi.org/10.1016/j.econlet.2026.113145).
- Anatolyev, S. and M. Sølvesten (2023). “Testing many restrictions under heteroskedasticity”. *Journal of Econometrics* 236.1, 105473. DOI: [10.1016/j.jeconom.2023.03.011](https://doi.org/10.1016/j.jeconom.2023.03.011).
- Boot, T. (2023). “Joint inference based on Stein-type averaging estimators in the linear regression model”. *Journal of Econometrics* 235.2, 1542–1563. DOI: [10.1016/j.jeconom.2023.01.006](https://doi.org/10.1016/j.jeconom.2023.01.006).
- Cattaneo, M. D., M. Jansson, and W. K. Newey (2018). “Inference in linear regression models with many covariates and heteroskedasticity”. *Journal of the American Statistical Association* 113.523, 1350–1361. DOI: [10.1080/01621459.2017.1328360](https://doi.org/10.1080/01621459.2017.1328360).
- Chao, J. C., J. A. Hausman, W. K. Newey, N. R. Swanson, and T. Woutersen (2014). “Testing overidentifying restrictions with many instruments and heteroskedasticity”. *Journal of Econometrics* 178, 15–21. DOI: [10.1016/j.jeconom.2013.08.003](https://doi.org/10.1016/j.jeconom.2013.08.003).
- Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012). “Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments”. *Econometric Theory* 28.1, 42–86. DOI: [10.1017/S0266466611000120](https://doi.org/10.1017/S0266466611000120).
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. K. Newey, and J. M. Robins (2018). “Double/debiased machine learning for treatment and structural parameters”. *Econometrics Journal* 21.1, C1–C68. DOI: [10.1111/ectj.12097](https://doi.org/10.1111/ectj.12097).
- Donald, S. G., G. W. Imbens, and W. K. Newey (2003). “Empirical likelihood estimation and consistent tests with conditional moment restrictions”. *Journal of Econometrics* 117.1, 55–93. DOI: [10.1016/S0304-4076\(03\)00118-0](https://doi.org/10.1016/S0304-4076(03)00118-0).
- Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, and K. Baicker (2012). “The Oregon Health Insurance Experiment: evidence from the first year”. *Quarterly Journal of Economics* 127.3, and the Oregon Health Study Group, 1057–1106. DOI: [10.1093/qje/qjs020](https://doi.org/10.1093/qje/qjs020).
- Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012). “Instrumental variable estimation with heteroskedasticity and many instruments”. *Quantitative Economics* 3.2, 211–255. DOI: [10.3982/QE89](https://doi.org/10.3982/QE89).
- Hong, Y. and H. White (1995). “Consistent specification testing via nonparametric series regression”. *Econometrica* 63.5, 1133–1159. DOI: [10.2307/2171724](https://doi.org/10.2307/2171724).

- Jochmans, K. (2022). “Heteroscedasticity-robust inference in linear regression models with many covariates”. *Journal of the American Statistical Association* 117.538, 887–896. DOI: [10.1080/01621459.2020.1831924](#).
- Kline, P., R. Saggio, and M. Sølvesten (2020). “Leave-out estimation of variance components”. *Econometrica* 88.5, 1859–1898. DOI: [10.3982/ECTA16410](#).
- Kolesár, M., H. Min, Y. Wang, and S. Zhang (2026). “Cluster-robust inference for quadratic forms”. arXiv: [2602.13537](#).
- Lee, Y. and R. Okui (2012). “Hahn–Hausman test as a specification test”. *Journal of Econometrics* 167.1, 133–139. DOI: [10.1016/j.jeconom.2011.10.005](#).
- Mikusheva, A. and L. Sun (2022). “Inference with many weak instruments”. *Review of Economic Studies* 89.5, 2663–2686. DOI: [10.1093/restud/rdab097](#).
- Newey, W. K. (1997). “Convergence rates and asymptotic normality for series estimators”. *Journal of Econometrics* 79.1, 147–168. DOI: [10.1016/S0304-4076\(97\)00011-0](#).
- Newey, W. K. and J. R. Robins (2018). “Cross-fitting and fast remainder rates for semiparametric estimation”. arXiv: [1801.09138](#).
- Sun, Y. and Q. Li (2006). “An alternative series based consistent model specification test”. *Economics Letters* 93.1, 37–44. DOI: [10.1016/j.econlet.2006.03.038](#).
- Tripathi, G. and Y. Kitamura (2003). “Testing conditional moment restrictions”. *Annals of Statistics* 31.6, 2059–2095. DOI: [10.1214/aos/1074290337](#).