Semiparametrically Efficient Inference for Kernel Measures of Noise Heterogeneity

Jakub Wornbard[*]
Gatsby Computational Neuroscience Unit
University College London

Zikai Shen[*]
Department of Statistical Science
University College London

Dimitri Meunier
Gatsby Computational Neuroscience Unit
University College London

Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London

Abstract

We develop semiparametrically efficient inference for kernel measures of noise heterogeneity in additive noise models. In many applications, the regression function is estimated using flexible machine learning methods. Downstream procedures based on the resulting residuals can then inherit first-stage bias: regression error may induce spurious dependence between covariates and residuals, invalidating the assumptions needed for standard analysis. We construct a novel Hilbert-valued one-step estimator of the kernel covariance operator between covariates and residuals. Our estimator yields bootstrap-calibrated tests for residual independence and goodness of fit in additive noise models, while also providing asymptotically efficient confidence intervals for the kernel dependence measure under noise heterogeneity. Our method also allows adjusting for additional covariates, enabling inference on distributional heterogeneity of residual noise across treatment groups. Simulations show improved power relative to naive plug-in residual methods, while remaining asymptotically calibrated.

1 Introduction

Many statistical and machine learning workflows use residuals to assess whether a fitted model has captured all systematic structure in the data. In additive noise models, this idea takes a particularly simple form: after regressing an outcome on covariates, the remaining noise should be independent of the variables whose effect has been explained. More broadly, one may ask whether the distribution of residual noise is homogeneous across covariates, treatment groups, or other features of interest. Kernel dependence measures provide a flexible way to quantify such residual noise heterogeneity, since they can detect general distributional differences beyond changes in mean or variance.

Classical nonparametric independence testing assumes that the variables to be tested are directly observed [Gretton et al., 2007, Heller et al., 2013, Berrett and Samworth, 2019, Shekhar et al., 2023]. In residual-based problems, however, the noise is not observed directly. It is generated by a first-stage regression estimator, often fit using flexible machine learning methods. This creates a multi-stage inference problem: downstream kernel procedures may inherit bias from the first-stage nuisance estimator. In this work, we bridge two strands of the statistical inference literature. On the one hand, recent work combines machine learning with semiparametric efficiency theory to estimate finite-dimensional target parameters [Chernozhukov et al., 2018, Hines et al., 2022, Kennedy, 2024]. On

[*]Equal contribution. Correspondence to Jakub Wornbard: jakub.wornbard.24@ucl.ac.uk.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the other hand, reproducing kernel Hilbert space (RKHS)-based dependence measures have become standard tools for nonparametric dependence testing [Gretton et al., 2005, 2007, Chwialkowski et al., 2014, Zhang et al., 2018, Albert et al., 2022, Shekhar et al., 2023] and related tasks in machine learning [Song et al., 2012, Shen et al., 2009, Peters et al., 2014, Li et al., 2021]. This raises the question of whether semiparametric debiasing can be used to construct valid kernel inference procedures when the variables entering the kernel statistic are themselves estimated.

To make the exposition concrete, we begin with the additive noise model studied by Peters et al. [2014] in the context of causal discovery. The data consists of n i.i.d. pairs $(x_i, y_i)_{i=1}^n$ drawn from the same distribution as $(X, Y) \sim P$. We consider the composite null hypothesis that there exist a function m, where we do not specify a parametric model for m, and a mean-zero noise variable ϵ independent of X such that $Y = m(X) + \epsilon$. If the true residuals $\epsilon_i = y_i - m(x_i)$ were observed, goodness of fit could be assessed by testing whether X and ϵ are independent. In practice, m is unknown, so an investigator estimates it by $\widehat{m}$ and constructs residuals $\widehat{\epsilon}_i = y_i - \widehat{m}(x_i)$. A natural plug-in approach is then to apply an off-the-shelf nonparametric independence test to the sample $\{(x_i, \widehat{\epsilon}_i)\}_{i=1}^n$.

We focus on testing using an RKHS with characteristic kernels [Sriperumbudur et al., 2010, 2011], which are state-of-the-art in nonparametric testing due to their ease of implementation and ability to be powerful against any deviation from the null, without specifying a parametric form. Specifically, we work with dependence measures between two probability distributions leveraging a kernelized cross-covariance operator. Our problem is a special case of the classical Hilbert–Schmidt Independence Criterion (HSIC) [Gretton et al., 2005], which has been widely deployed in practice to assess independence of two random variables, as applied to a covariate sample and an estimated residual sample. The challenge, and our contribution, lies in extending HSIC to the challenging case with estimated residuals, and obtaining a calibrated and powerful test for detecting residual noise heterogeneity.

We now illustrate the challenge to obtain calibrated and powerful tests with plug-in estimators of the residuals. Indeed, the estimated residuals $\widehat{\epsilon}_i = \epsilon_i + m(x_i) - \widehat{m}(x_i)$ can induce spurious dependence between X_i and $\widehat{\epsilon}_i$ under the composite null hypothesis. As a simple illustration of this challenge, suppose we split the data into two disjoint folds: one fold is used to estimate the regression function m, and the other fold is used to construct the independence test. If the fitting fold is small, $\widehat{m}$ may have substantial bias, and the kernel test may interpret this first-stage bias as evidence of dependence, leading to inflated Type I error. If the fitting fold is large, the residuals better approximate the oracle residuals, but the holdout sample used for testing is small, leading to low power in practice. Related empirical failures have been observed in prior work on conditional independence testing [Pogodin et al., 2024, Wang et al., 2025, He et al., 2025]. Sample splitting removes the direct dependence between the fitted regression algorithm and the test sample, but conditional on the fitted function the residual still contains the systematic error $m(X) - \widehat{m}(X)$, which is a function of X. Our discussion highlights the practical difficulty of controlling Type I error while retaining power for tests with estimated nuisance functions. Recent work such as Key et al. [2025], Huang and Niu [2026] on kernel testing with composite null addresses the setting where finite-dimensional nuisance parameters are estimated, which do not provide guarantees for our setting with nonparametric nuisances.

Our approach is based on a debiasing device to remove the first-order bias of estimated regressions, making use of the efficient influence function of a target functional [Bickel et al., 1998, Kennedy, 2024]. Nevertheless there is a freedom in which object we choose to debias. While it appears natural to attempt to directly debias the canonical scalar-valued HSIC test statistic, this presents a complication: the empirical unbiased HSIC is a degenerate U-statistic under the null [Lee, 1990]. At the population level, this first-order degeneracy appears as a vanishing efficient influence function: the first-order efficient influence function of

$$\mathrm{HSIC}(X, Y - \mathbb{E}[Y \mid X])$$

is identically zero under the null. An even more immediate way to see this issue is to note that HSIC is the squared RKHS norm of the cross-covariance operator, and the degeneracy is a manifestation of the fact that the derivative of $x \mapsto \|x\|^2$ vanishes at zero. Therefore, first-order scalar debiasing of HSIC cannot remove the first-stage bias that matters for obtaining a calibrated test at the null.

Noting that the issue is caused by taking the squared norm, we instead use a Hilbert-valued efficient influence function to debias the cross covariance operator. The resultant construction, called a one-step estimator, satisfies a Hilbert-valued central limit theorem, yielding the crucial distributional guarantees we require to obtain a calibrated test. Our work builds on work on Hilbert-valued semiparametric efficiency theory [Luedtke and Chung, 2024, Morzywolek et al., 2025]. We obtain a scalar kernel dependence measure by taking the squared Hilbert–Schmidt norm of the debiased estimator. The resulting test is calibrated under the null, where scalar HSIC is degenerate, and the confidence intervals for the kernel dependence measure are efficient under fixed alternatives. Related doubly-robust kernel tests have been developed for distributional treatment effects [Fawkes et al., 2024, Martinez-Taboada et al., 2023].

Contributions. We propose a Hilbert-valued one-step estimator for the kernel covariance operator between covariates and residuals, whose squared norm gives the corresponding kernel dependence measure. We prove asymptotic linearity and weak convergence of the estimator under nonparametric nuisance estimation. These results yield a bootstrap-calibrated independence test with asymptotic validity under the null and consistency against fixed alternatives. Under alternatives, our approach also provides asymptotically efficient confidence intervals for the value of the kernel dependence measure. Our framework is highly flexible and also straightforwardly permits adjustment for additional covariates, which is a natural application for diverse treatment groups. Finally, we conduct extensive simulations showing that the proposed method improves power relative to naive plug-in residual methods, while importantly remaining calibrated.

Outline. In Section 2, we introduce the problem setup and the RKHS covariance operator target. In Section 3, we derive the Hilbert-valued efficient influence function, define the one-step estimator, and establish asymptotic linearity, weak convergence, and bootstrap calibration. Section 4 presents synthetic experiments illustrating calibration, power, and confidence-interval coverage. Section 5 discusses limitations and future directions.

2 Problem Setup

We study independence testing between random variables X and Y, after adjusting for a covariate, denoted as W. We apply our method to two practical settings, namely goodness-of-fit testing for additive noise models and testing for noise homogeneity across different treatment groups, see Section 4. We refer to these as ***Settings 1*** and ***2*** respectively. In the former case, $X = W$, and our data consists of (X, Y)-pairs. In the latter case, X is a treatment indicator, and $W = (X, T)$ where T are additional covariates. We do not test for homogeneity with respect to the covariates T. For the remainder of the paper, we make the following assumption, although our theory readily generalizes to the case where X is not measurable with respect to the covariate W we adjust on.

Assumption 1. *There exists a measurable map $\pi : \mathcal{W} \to \mathcal{X}$ such that $X = \pi(W)$.*

Let $\mathcal{O} = \mathcal{W} \times \mathbb{R}^{d_y}$, and let $\mathcal{M}$ denote a nonparametric model of probability laws on $\mathcal{O}$. We write $P_0 \in \mathcal{M}$ for the true data-generating distribution. For $P \in \mathcal{M}$, define $m_P(w) \doteq \mathbb{E}_P[Y \mid W = w]$ and $\xi_P(w, y) \doteq y - m_P(w)$. We test $H_0 : P_0 \in \{P \in \mathcal{M} : \xi_P(W, Y) \perp X\}$. In ***Setting 2***, testing for H_0 addresses the question of whether the marginal distribution of the residual is independent of X, after adjusting for X and additional covariates T.

We next introduce necessary RKHS concepts. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric positive definite kernel and let $\mathcal{H}_k$ be its reproducing kernel Hilbert space (RKHS) [Berlinet and Thomas-Agnan, 2011]. We denote the canonical feature map of $\mathcal{H}_k$ as $\varphi_k(x) = k(\cdot, x)$. Let $l : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ be a symmetric positive definite kernel with RKHS $\mathcal{H}_l$ and feature map $\varphi_l(u) = l(\cdot, u)$. We refer the reader to Appendix B for additional background on RKHSs. We assume all RKHSs mentioned are separable. This is satisfied, e.g. if the reproducing kernel is continuous and the domain is separable.

Throughout, we assume the kernels are measurable and satisfy the moment conditions needed for the following Bochner expectations to exist; for example, the following assumption is sufficient.

Assumption 2 (Kernel boundedness)**.** $\sup_{x \in \mathcal{X}} k(x, x) \leq \kappa_k^2$ *and* $\sup_{u \in \mathbb{R}^{d_y}} l(u, u) \leq \kappa_l^2$.

The dependence between X and the residual $\xi_P(W,Y)$ is encoded by first featurizing them with infinite dimensional kernel features and then taking the expectation of their outer product. In particular, we use the language of kernel mean embeddings [Smola et al., 2007]. We denote the kernel mean embeddings of X and $\xi_P(W,Y)$ as $\mu_{X,P} \doteq \mathbb{E}_P[\varphi_k(X)] \in \mathcal{H}_k$ and $\mu_{\xi,P} \doteq \mathbb{E}_P[\varphi_l(\xi_P(W,Y))] \in \mathcal{H}_l$. For characteristic kernels, testing for dependence between X and $\xi_P(W,Y)$ nonparametrically is equivalent to testing whether the kernel mean embedding of the tensor product feature map $\mathbb{E}_P[\varphi_k(X)\otimes\varphi_l(\xi_P(W,Y))]$ is equal to $\mu_{X,P}\otimes\mu_{\xi,P}$ [Gretton, 2015]. This is algebraically equivalent to testing for $\Psi(P)=0$, where

$$\Psi(P) = \mathbb{E}_P\left[\widetilde{\varphi}_{k,P}(X) \otimes \widetilde{\varphi}_{l,P}(\xi_P(W,Y))\right] \in \mathcal{H}_{kl},$$

where we define the *centered* feature maps $\widetilde{\varphi}_{k,P}(x) \doteq \varphi_k(x) - \mu_{X,P}$ and $\widetilde{\varphi}_{l,P}(\xi_P(w,y)) \doteq \varphi_l(\xi_P(w,y)) - \mu_{\xi,P}$. $\mathcal{H}_{kl} \doteq \mathcal{H}_k \otimes \mathcal{H}_l$ denotes the tensor product RKHS induced by the product kernel $(x,u),(x',u') \mapsto k(x,x')l(u,u')$. $\|\Psi(P)\|^2_{\mathcal{H}_{kl}}$ is the Hilbert–Schmidt Independence Criterion (HSIC) between X and $Y - m_P(W)$ [Gretton et al., 2005]. We define the hypothesis test

$$H_0 : \Psi(P_0) = 0 \qquad \text{against} \qquad H_1 : \Psi(P_0) \neq 0.$$

We provide additional background on HSIC independence tests and the interpretation of their results in Appendix L. Throughout, a subscript 0 denotes evaluation at the true distribution P_0; for example, $m_0 := m_{P_0}$ and $\Psi_0 := \Psi(P_0)$.

3 One-step Estimator

Semiparametric efficiency theory offers an attractive possibility of debiased two-stage learning, by computing an efficient influence function (EIF) and then removing the first order bias of the first stage estimator. However, the scalar-valued HSIC is first-order degenerate at the null H_0 as its EIF at the null is identically zero. The correct object to debias is thus the Hilbert-valued cross-covariance operator $\Psi(P)$, and we should compute the squared norm ***after*** debiasing. We now derive the EIF of the cross-covariance. Let $L^2_0(P_0)$ denote the space of square-integrable functions with zero mean under P_0.

Assumption 3 (Model and moment conditions)**.** *Let $\mathcal{M}$ be a nonparametric model with tangent space $\dot{\mathcal{M}}_{P_0} = L^2_0(P_0)$. The laws in $\mathcal{M}$ are mutually absolutely continuous and satisfy $\sup_{P\in\mathcal{M}} \|\mathbb{E}_P[\|Y\|^2 \mid W]\|_{L^\infty(P_W)} < \infty$. We assume that $\mathbb{E}_0\|Y\|^2 < \infty$.*

Assumption 4 (Regularity of the residual kernel)**.** *The canonical feature map $\varphi_l : \mathbb{R}^{d_y} \to \mathcal{H}_l$ of the symmetric positive definite kernel $l : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is continuously Fréchet differentiable. For some finite constant $L_2 \geq 0$, for $u, v \in \mathbb{R}^{d_y}$,*

$$\|D\varphi_l(u) - D\varphi_l(v)\|_{\mathrm{op}} \leq L_2\|u - v\|.$$

By Lemma 1, Assumptions 2 and 4 imply, for all $u, h \in \mathbb{R}^{d_y}$, $\sup_v \|D\varphi_l(v)\|_{\mathrm{op}} \leq \sqrt{2\kappa_l L_2}$ and

$$\|\varphi_l(u+h) - \varphi_l(u) - D\varphi_l(u)[h]\|_{\mathcal{H}_l} \leq \frac{L_2}{2}\|h\|^2. \tag{1}$$

We give sufficient conditions in terms of derivatives of l in Lemma 2 and verify them for Gaussian and Sobolev kernels in Proposition 1.

Intuition about the form of the EIF. The correction term in the EIF can be understood from a first-order expansion of the residual feature map. Let $\widehat{\delta}(w) \doteq \widehat{m}(w) - m_0(w)$. Then

$$\varphi_l\{Y - \widehat{m}(W)\} = \varphi_l\{\xi_0(W,Y) - \widehat{\delta}(W)\} \approx \varphi_l\{\xi_0(W,Y)\} - \sum_{j=1}^{d_y} \widehat{\delta}_j(W)\partial_j\varphi_l\{\xi_0(W,Y)\}.$$

Therefore, conditional on W, the first-order shift in the residual feature is

$$\text{first-order shift} = -\sum_{j=1}^{d_y} \widehat{\delta}_j(W)v_{0,j}(W), \qquad v_{0,j}(w) \doteq \mathbb{E}_0[\partial_j\varphi_l\{\xi_0(W,Y)\} \mid W = w].$$

This is the component of the plug-in error that can create spurious dependence with X. The EIF subtracts $\sum_{j=1}^{d_y} \xi_{0,j}(W,Y)v_{0,j}(W)$. For the plug-in residual $\widehat{\xi}_j = \xi_{0,j} - \widehat{\delta}_j(W)$, this correction changes by $+\sum_{j=1}^{d_y} \widehat{\delta}_j(W)v_{0,j}(W)$, which cancels the conditional first-order Taylor bias above. Equivalently, after the correction, the leading perturbation is proportional to $\sum_{j=1}^{d_y} \widehat{\delta}_j(W)\left[\partial_j\varphi_l\{\xi_0(W,Y)\} - v_{0,j}(W)\right]$, whose conditional mean given W is zero. Thus the first-order effect of regression error is removed, leaving only higher-order remainder terms.

Theorem 1 (Efficient influence function)**.** *Suppose Assumptions 1 to 4 hold. Then $\Psi : \mathcal{M} \to \mathcal{H}_{kl}$ is pathwise differentiable at P_0. For any regular parametric submodel $t \mapsto P_t$ through P_0 with score $s \in L_0^2(P_0)$, $\dot{\Psi}_0(s) \doteq \lim_{t\to 0} \frac{1}{t}(\Psi(P_t) - \Psi(P_0)) = \mathbb{E}_0[\phi_0(O)s(O)] \in \mathcal{H}_{kl}$, where for $o = (x, w, y)$,*

$$\phi_0(o) \doteq \widetilde{\varphi}_{k,0}(x) \otimes \left[\widetilde{\varphi}_{l,0}(\xi_0(w,y)) - \sum_{j=1}^{d_y} \xi_{0,j}(w,y)v_{0,j}(w)\right] - \Psi_0.$$

ϕ_0 belongs to $L_0^2(P_0; \mathcal{H}_{kl})$, therefore it is the $\mathcal{H}_{kl}$-valued EIF of Ψ at P_0 in the sense of Luedtke and Chung [2024].

A slightly more general form of Theorem 1 is proved in Appendix C.

3.1 One-step estimator and asymptotic normality

We now use the EIF to perform one-step debiasing for Ψ_0. We refer to the debiased estimator as a one-step estimator [Schick, 1986, Luedtke and Chung, 2024]. We will establish Hilbert-valued asymptotic normality of this estimator under a set of reasonable assumptions. We may use the asymptotic normality theorem to construct a confidence interval for any scalar-valued transformation of Ψ_0. We choose the RKHS norm of Ψ_0, which is sufficient for testing for whether $\Psi_0 = 0$. A hypothesis test with Type I error control directly leads to a confidence interval for the true value Ψ_0. We refer to such a test as calibrated. By including all values of Ψ_0 not rejected by a test with Type I error control, one obtains a confidence set with the desired coverage level. We use this duality to obtain a confidence interval for $\|\Psi_0\|_{\mathcal{H}_{kl}}$ and a hypothesis test for H_0.

We assume the dataset is $\{o_i = (x_i, w_i, y_i)\}_{i=1}^n$, where n is even. Let $I_1 = \{1, \ldots, n/2\}$ and $I_2 = \{n/2 + 1, \ldots, n\}$. We fit the nuisance functions on I_2 and evaluate the one-step estimator on I_1. Write $\widehat{\xi}_i \doteq y_i - \widehat{m}(w_i)$, where $\widehat{m}$ is *any* machine learning estimator for m_0. Define the empirical kernel mean embeddings $\widehat{\mu}_X \doteq \frac{2}{n}\sum_{i\in I_2} \varphi_k(x_i)$ and $\widehat{\mu}_\xi \doteq \frac{2}{n}\sum_{i\in I_2} \varphi_l(\widehat{\xi}_i)$.

For $j = 1, \ldots, d_y$, we estimate the $\mathcal{H}_l$-valued nuisance $v_{0,j}$ by vector-valued kernel ridge regression [Caponnetto and De Vito, 2007, Park and Muandet, 2020, Li et al., 2024]. Let $\mathcal{H}_{ql}$ denote the vector-valued RKHS induced by the kernel $q(\cdot,\cdot)\mathrm{Id}_{\mathcal{H}_l}$. For details on vector-valued kernels and their associated RKHS, we refer the reader to Li et al. [2024] and Appendix E. We define, for each $j = 1, \ldots, d_y$,

$$\widehat{v}_j = \operatorname*{arg\,min}_{v\in\mathcal{H}_{ql}} \frac{2}{n}\sum_{i\in I_2} \left\|(\partial_j\varphi_l)(\widehat{\xi}_i) - v(w_i)\right\|_{\mathcal{H}_l}^2 + \lambda_{j,n}\|v\|_{\mathcal{H}_{ql}}^2.$$

For $i \in I_1$, define $\widehat{M}_i \doteq \varphi_l(\widehat{\xi}_i) - \widehat{\mu}_\xi - \sum_{j=1}^{d_y} \widehat{\xi}_{i,j}\widehat{v}_j(w_i)$. The one-step estimator evaluated on I_1 is

$$\widehat{\Psi}_n \doteq \frac{2}{n}\sum_{i\in I_1} (\varphi_k(x_i) - \widehat{\mu}_X) \otimes \widehat{M}_i. \tag{2}$$

Remark 1 (Cross-fitting)**.** *To simplify our exposition and notations, in the main document we work exclusively with a sample split estimator $\widehat{\Psi}_n$, introduced above in Eq. (2). Since $\widehat{\Psi}_n$ is constructed by evaluating half the samples, it leads to a loss of statistical power. To remedy this, we construct a K-fold cross-fitted estimator, where we use all K folds except one for fitting, one fold for evaluation, and average over all K choices of the evaluation fold. We emphasize that all simulation experiments and theoretical guarantees in our work apply to the K-fold cross-fitted estimator, as is done in the*

*Appendix. The cross-fitted estimator allows us to remove restrictive Donsker-type conditions on the function class [Schick, 1986, Chernozhukov et al., 2018]. We provide generalized results for the cross-fitted estimator in the Appendix, denoted with **reverse hat** notation as $\check{\Psi}_n$. We develop the core asymptotic results for those in Appendix C assuming $K = 2$. Obtaining analogues for higher values is not harder but requires more careful notation. This simplification is standard in the literature [Chernozhukov et al., 2018].*

Requirement 1 (Conditions on the nuisance estimators)**.**

(A) For a finite constant $C > 0$, we have $\|m_0\|_{L^\infty(P_{0,W};\mathbb{R}^{d_y})} \leq C$ and

$$\operatorname*{ess\,sup}_{w \in \operatorname{supp}(P_{0,W})} \mathbb{E}_0[\|Y - m_0(W)\|^2_{\mathbb{R}^{d_y}} \mid W = w] \leq C. \tag{3}$$

(B) We have

$$\max\left\{\|\widehat{m} - m_0\|_{L^\infty(P_{0,W};\mathbb{R}^{d_y})}, \|\widehat{m} - m_0\|_{L^2(P_{0,W};\mathbb{R}^{d_y})}\right\} = o_p(1)$$
$$\max\left\{\|\widehat{v} - v_0\|_{L^2(P_{0,W};\mathcal{H}_l^{d_y})}, \|\widehat{\mu}_\xi - \mu_{\xi,0}\|_{\mathcal{H}_l}\right\} = o_p(1).$$

(C) We have $\|\widehat{m} - m_0\|^2_{L^2(P_{0,W};\mathbb{R}^{d_y})} + \|\widehat{m} - m_0\|_{L^2(P_{0,W};\mathbb{R}^{d_y})}\|\widehat{v} - v_0\|_{L^2(P_{0,W};\mathcal{H}_l^{d_y})} = o_p(n^{-1/2})$.

Note that we do ***not*** have the mixed bias property [Rotnitzky et al., 2021] with respect to the nuisances m_0 and v_0, since we always require $\widehat{m} - m_0$ to converge at $n^{-1/4}$-rate.

Theorem 2 (Asymptotic linearity and weak convergence of the sample-split one-step estimator)**.** *Suppose the conclusion of Theorem 1 holds and let ϕ_0 be given as in Theorem 1. Under Assumption 4 and Requirement 1, the sample-split one-step estimator $\widehat{\Psi}_n$ is regular and asymptotically linear:*

$$\sqrt{n/2}\left(\widehat{\Psi}_n - \Psi_0\right) = \sqrt{\frac{2}{n}}\sum_{i \in I_1} \phi_0(o_i) + o_p(1) \rightsquigarrow \mathbb{H}, \tag{4}$$

where $\mathbb{H}$ is a tight $\mathcal{H}_{kl}$-valued Gaussian random variable such that, for each $h \in \mathcal{H}_{kl}$, $\langle \mathbb{H}, h\rangle_{\mathcal{H}_{kl}} \sim \mathcal{N}(0, \mathbb{E}_0[\langle \phi_0(X, W, Y), h\rangle^2_{\mathcal{H}_{kl}}])$.

When is Requirement 1 satisfied? Requirements 1B and 1C are a set of sufficient conditions that hold under more fundamental assumptions, such as membership in a Hölder smoothness class. $\widehat{m}$ is a standard nonparametric regression estimator for m_0. Its minimax rate of convergence is well understood [Tsybakov, 2009]. Both $\widehat{v}$ and $\widehat{\mu}_\xi$ depend on the nonparametric first stage $\widehat{m}$. In the second stage, $\widehat{v}$ is obtained from performing a vector-valued kernel ridge regression, while $\widehat{\mu}_\xi$ is obtained from computing a sample average. In the former case, the convergence rates depend on the smoothness of the vector-valued target v_0 and the capacity of the vector-valued RKHS $\mathcal{H}_{ql}$. In the latter case, the second stage error is a Monte Carlo error. We emphasize that thanks to the second order nature of the nuisance rates in Requirement 1C, nowhere do we require a nuisance function to converge at the parametric rate $n^{-\frac{1}{2}}$. In Appendix E, we provide the necessary background on vector-valued kernel learning, state and prove technical theorems on the convergence rates of $\widehat{v}$ and $\widehat{\mu}_\xi$, and provide a concrete example where Requirement 1 is satisfied.

3.2 Debiased confidence set and hypothesis test

We now construct confidence intervals for $\|\Psi_0\|_{\mathcal{H}_{kl}} = \sqrt{\text{HSIC}}$ and $\|\Psi_0\|^2_{\mathcal{H}_{kl}} = \text{HSIC}$. We first obtain a bootstrap interval for the norm and a test of $H_0 : \Psi_0 = 0$, then derive efficient delta-method intervals on both scales under fixed alternatives. Both constructions rely on the asymptotic normality result of Theorem 2.

By the continuous mapping theorem applied to Eq. (4), we deduce that $(n/2)\|\widehat{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}} \rightsquigarrow \|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$. We can approximate $\|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$ by the bootstrapped distribution of $(n/2)\|\widehat{\Psi}_n - \widehat{\Psi}_n^{(b)}\|^2_{\mathcal{H}_{kl}}$

[Efron, 1992]. We briefly sketch our procedure, and refer the reader to Appendix G for a formal exposition. For $b \in [B]$, we let $o_1^{(b)}, \dots, o_{\frac{n}{2}}^{(b)} \overset{\text{i.i.d.}}{\sim} \frac{2}{n} \sum_{i \in I_1} \delta_{o_i}$, conditionally on the original data. Let ζ_n denote the $(1-\alpha)$-quantile of $\left\| \sqrt{n/2}(\widehat{\Psi}_n - \widehat{\Psi}_n^{(b)}) \right\|_{\mathcal{H}_{kl}}^2$, where $\widehat{\Psi}_n^{(b)}$ is given as in Eq. (2) except evaluated on the bth bootstrap sample. Then ζ_n converges to the $(1-\alpha)$-quantile $\zeta_{1-\alpha}$ of $\|\mathbb{H}\|_{\mathcal{H}_{kl}}^2$, provided the estimated EIF is a consistent estimator for the true EIF. We refer the reader to Luedtke and Chung [2024, Theorem 4] for the proof. In simulations, we replace ζ_n by its finite-B approximation $\widehat{\zeta}_n$.

Armed with a consistent estimate ζ_n for the $(1-\alpha)$-quantile of $\|\mathbb{H}\|_{\mathcal{H}_{kl}}^2$, we propose the following confidence interval for $\|\Psi_0\|_{\mathcal{H}_{kl}}$:

$$\mathcal{I}_{\zeta_n} \doteq \left[\left\{ \|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}} - \sqrt{2\zeta_n/n} \right\}_+, \ \|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}} + \sqrt{2\zeta_n/n} \right]. \tag{5}$$

The reverse triangular inequality gives $\left| \|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}} - \|\Psi_0\|_{\mathcal{H}_{kl}} \right| \leq \|\widehat{\Psi}_n - \Psi_0\|_{\mathcal{H}_{kl}}$. Thus $\mathcal{I}_{\zeta_n}$ contains $\|\Psi_0\|_{\mathcal{H}_{kl}}$ whenever $\|\widehat{\Psi}_n - \Psi_0\|_{\mathcal{H}_{kl}} \leq \sqrt{2\zeta_n/n}$. At $\Psi_0 = 0$, the reverse triangular inequality is an equality. We reject H_0 if and only if $0 \notin \mathcal{I}_{\zeta_n}$.

For fixed $\Psi_0 \neq 0$, we may also construct asymptotically efficient normal confidence intervals.

$$\begin{aligned} \sqrt{n/2} \left(\|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}}^2 - \|\Psi_0\|_{\mathcal{H}_{kl}}^2 \right) &= \sqrt{n/2} \left\langle \widehat{\Psi}_n - \Psi_0, \widehat{\Psi}_n + \Psi_0 \right\rangle_{\mathcal{H}_{kl}} \\ &= \sqrt{n/2} \|\widehat{\Psi}_n - \Psi_0\|_{\mathcal{H}_{kl}}^2 + \sqrt{2n} \langle \widehat{\Psi}_n - \Psi_0, \Psi_0 \rangle_{\mathcal{H}_{kl}}. \end{aligned} \tag{6}$$

The second term is asymptotically normal with mean zero with asymptotic variance $\mathbb{E}_0[4\langle \phi_0(X,W,Y), \Psi_0 \rangle_{\mathcal{H}_{kl}}^2]$, whereas the first term is $O_p(1/\sqrt{n})$. As a result, we have a CLT for $\|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}}^2$, namely $\sqrt{n/2} \left(\|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}}^2 - \|\Psi_0\|_{\mathcal{H}_{kl}}^2 \right) \rightsquigarrow \mathcal{N}(0, \mathbb{E}_0[4\langle \phi_0(X,W,Y), \Psi_0 \rangle_{\mathcal{H}_{kl}}^2])$. The limiting variance vanishes at $\Psi_0 = 0$, which is precisely what prevents us from debiasing the scalar-valued HSIC directly. When $\sigma^2 \doteq \mathbb{E}_0[\langle \phi_0(X,W,Y), \Psi_0 \rangle_{\mathcal{H}_{kl}}^2]$ is positive and consistently estimated by $\widehat{\sigma}_n^2$, we define the following confidence interval for HSIC:

$$\mathcal{I}_\delta^{1-\alpha} \doteq \left[\|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}}^2 - 2z_{1-\alpha/2} \sqrt{\frac{2\widehat{\sigma}_n^2}{n}}, \ \|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}}^2 + 2z_{1-\alpha/2} \sqrt{\frac{2\widehat{\sigma}_n^2}{n}} \right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution. From van der Vaart [2000, Theorem 25.47], we know that $\mathbb{E}_0[4\langle \phi_0(X,W,Y), \Psi_0 \rangle_{\mathcal{H}_{kl}}^2]$ is the semiparametric efficiency bound for estimating $\|\Psi_0\|_{\mathcal{H}_{kl}}^2$ with $\frac{n}{2}$ samples, which suggests that when $\Psi_0 \neq 0$ holds, then $\mathcal{I}_\delta^{1-\alpha}$ may be much less wide than $\{t^2 : t \in \mathcal{I}_{\zeta_n}\}$ while being asymptotically calibrated. Applying the delta method once more to $t \mapsto \sqrt{t}$ gives efficient confidence intervals for $\sqrt{\text{HSIC}} = \|\Psi_0\|_{\mathcal{H}_{kl}}$, the scale used in Section 4. We establish the validity of this approach carefully in Theorem 4 of Appendix G.

4 Experiments

We verify the validity of the proposed procedures on a range of synthetic experiments, where we are able to explicitly control the extent to which noise homogeneity is violated, and to adjust the statistical difficulty of all the intermediate procedures.

Baselines. We compare against plug-in residual procedures that estimate m and then apply standard HSIC inference without the EIF correction. The split-sample baseline follows the ANM diagnostic of Peters et al. [2014]: fit $\hat{m}$ on a training split, compute residuals $Y_i - \hat{m}(W_i)$ on held-out observations, and run the permutation HSIC test [Gretton et al., 2007]. We report several training fractions because the split controls a calibration–power tradeoff.

We also consider a full-sample, cross-fitted plug-in permutation test. Let $I_1, \dots, I_K$ be the folds, and write $\hat{m}^{(-k)}$ for the regression estimator trained on all observations outside fold I_k. For $i \in I_k$,

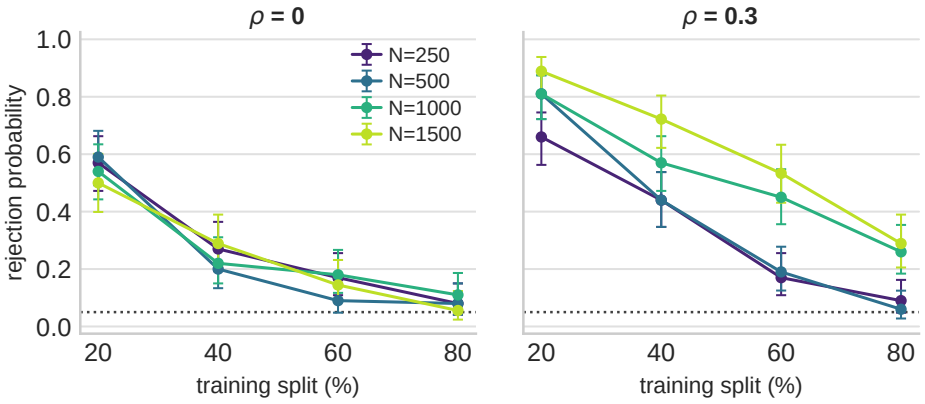


Figure 1: Split fit-test baseline at $s_m = 1, s_\epsilon = 0.25$.

define the out-of-fold residual $\hat{\xi}_i = Y_i - \hat{m}^{(-k)}(W_i)$, form Gram matrices $K_{ij} = k(X_i, X_j)$ and $L_{ij} = l(\hat{\xi}_i, \hat{\xi}_j)$, and perform a full-sample permutation test using these. This uses all observations and avoids in-sample residuals, but remains heuristic because the permutation null does not account for first-stage estimation error.

For confidence intervals, we also compare against plug-in methods, which use the same bootstrap or delta-method construction as our estimator, but with the EIF correction term omitted. This helps us evaluate the effect of bias correction on the performance.

X-dependent residuals. We first evaluate goodness-of-fit testing for the scalar additive-noise model $Y = m(X) + \epsilon$, $\epsilon \perp X$. Data are generated from

$$X \sim \mathrm{Unif}[-3, 3], \qquad Y = m(X) + \sigma(X)Z, \qquad Z \sim \mathcal{N}(0, 1).$$

Both m and $\log \sigma$ are sampled as truncated random Fourier series. The decay parameters s_m and s_ϵ control smoothness, and hence first-stage estimation difficulty: s_m is the exponent in the power-law decay rate of the Fourier coefficients of m, and s_ϵ plays an analogous role in $\log \sigma(X)$. ρ controls the strength of residual heterogeneity. Smaller values of those lead to rougher regression functions and less regular variance structure in the noise. ρ multiplies the non-constant X-dependent term in $\log \sigma(X)$: $\rho = 0$ gives homogeneous noise and larger ρ yields stronger X-dependence in the residual distribution. In each setting, we estimate $\sqrt{\mathrm{HSIC}(X, \epsilon)}$ using Gaussian kernels with bandwidths chosen by the median heuristic. We keep $d_x = d_y = 1$ in the reported experiments to isolate first-stage residual-estimation bias from the additional effects of high-dimensional kernel testing and nuisance estimation. The data generating mechanism is given in Appendix J.1.

Figure 2 reports rejection curves over ρ for a representative setting. At $\rho = 0$, the debiased bootstrap test is close to nominal, while the cross-fitted permutation baseline is conservative. As ρ increases, the debiased test gains power without requiring a split-ratio choice. So does the baseline, though it attains lower power. Figure 1 shows the split-sample plug-in baseline for the same synthetic residual experiment. Under the null, rejection should be near the nominal 5% line, but small training splits leave enough residual-estimation bias to inflate rejection. Under the alternative, larger training splits fit the regression better but leave fewer observations for the HSIC test, so power can fall sharply. This calibration–power tradeoff is the tuning problem that Figure 2 avoids: the debiased test uses the full sample and gains power without requiring a split-ratio choice.

Appendix J.7 compares plug-in and debiased estimators across kernel bandwidths more extensively. In several settings, debiasing increases operator MSE but reduces the MSE of the HSIC U-statistic. We note that these findings are compatible. The two squared-error losses are respectively $\|\widehat{\Psi} - \Psi_0\|^2$ and $(\widehat{\theta} - \theta)^2$, where $\theta = \|\Psi_0\|^2$ and $\widehat{\theta}$ denotes a U- or V-statistic estimator of θ.

Residual heterogeneity with additional covariates. We next consider a setting where the magnitude of Ψ_0, rather than only the null $\Psi_0 = 0$, is scientifically meaningful. Let $W = (X, T)$,

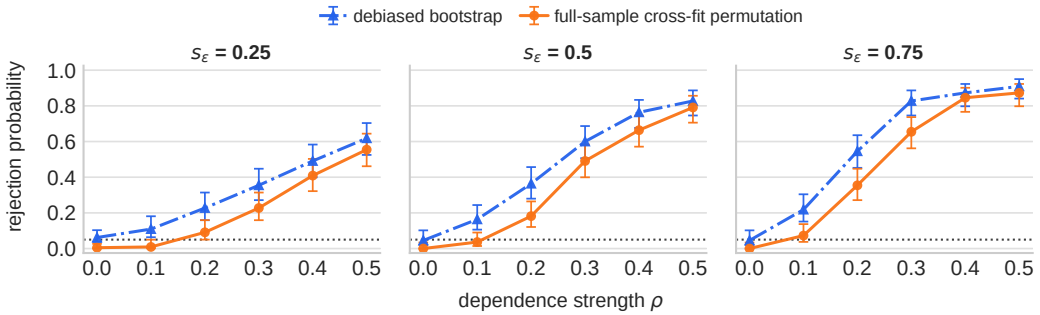


Figure 2: Rejection probabilities at $s_m = 1$ and $n = 500$ across noise-dependence levels and smoothness parameters s_ϵ of $\sigma(x)$, for the debiased confidence interval of Eq. (5) and the cross-fitted permutation test.

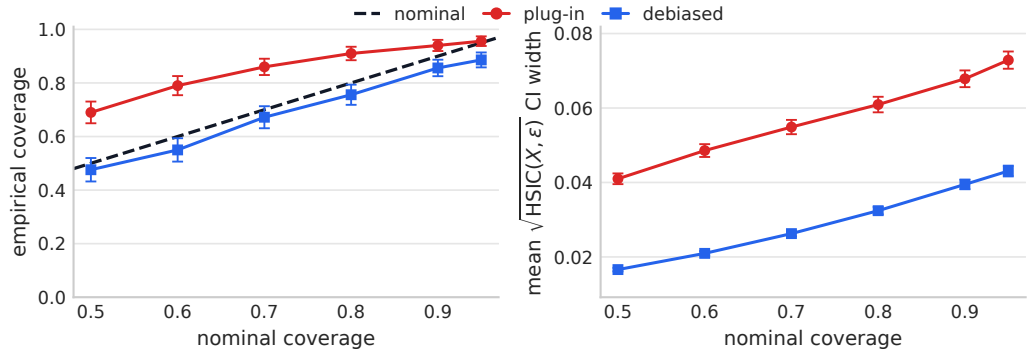


Figure 3: Calibration of delta-method confidence intervals for $\|\Psi_0\|_{\mathcal{H}_{kl}} = \sqrt{\text{HSIC}}$ when $W = (X, T)$.

where $X \in \{0, 1\}$ is a binary group indicator and T is an additional covariate. Writing $R = Y - \mathbb{E}[Y \mid X, T]$, the target measures whether the covariate-adjusted residual law differs across the two groups.

This connection is especially transparent with the discrete kernel $k(x, x') = \mathbf{1}\{x = x'\}$ on X. When the two groups are balanced,

$$\text{HSIC}(X, R) = \frac{1}{8}\text{MMD}_l^2(\mathcal{L}(R \mid X = 0), \mathcal{L}(R \mid X = 1)),$$

where MMD_l is the Maximum Mean Discrepancy [Gretton et al., 2007] associated with the residual kernel l, between the laws of the residuals under $X = 0$ and under $X = 1$. We evaluate this target in a synthetic model

$$Y = m(X, T) + \sigma_X Z, \qquad Z \sim \mathcal{N}(0, 1),$$

where m is a finite trigonometric series in T, and l is a fixed random-feature approximation to a Gaussian kernel. The null has equal residual scales across groups, while the alternative has $\sigma_0 \neq \sigma_1$. Full data-generating details are given in Appendix J.4.

Figure 3 shows that debiasing produces narrower confidence sets, though it also introduces some finite-sample undercoverage.

Causal discovery by two-way goodness-of-fit testing. Additive-noise causal discovery compares the two regression directions by fitting

$$Y = m_{Y|X}(X) + \epsilon_{Y|X} \qquad \text{and} \qquad X = m_{X|Y}(Y) + \epsilon_{X|Y},$$

and testing whether the fitted residuals are independent of the corresponding covariate [Peters et al., 2014]. In the ideal additive-noise setting, the causal direction should pass this goodness-of-fit test, while the reverse direction in general should fail, except for special non-identifiable cases such as linear m and Gaussian noise. This requires calibration under a fitted-residual null and power against a non-additive reverse direction.

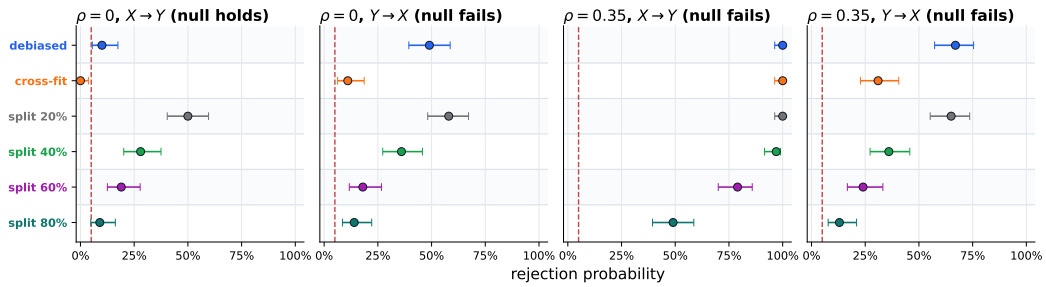


Figure 4: Rejection probabilities for nominal level-5% tests in the causal discovery experiment. "debiased" is the test based on Eq. (5); "cross-fit" is the proposed permutation baseline which utilises the entire sample and cross-fits the nuisances, and "split x%" refers to tests where x% of the sample is used for fitting the nuisances and the rest for a permutation test.

We study a synthetic bivariate example with known structural direction $X \to Y$. The structural function is nonlinear and resembles the low-dimensional cause-effect pairs used in the Tübingen cause-effect pair benchmark [Mooij et al., 2016]. We consider two regimes. In the first, the forward model has homogeneous additive noise, so the ANM null holds for the regression of Y on X and fails in the reverse direction. In the second, the structural direction remains $X \to Y$, but the forward residual law depends on X; in this case both ANM goodness-of-fit nulls should be rejected. Full data-generating details and example datasets are given in Appendix J.3. Figure 4 shows that the debiased bootstrap gives the most favourable two-way goodness-of-fit profile, and highlights the failure mode of split-sample tests. The latter only become calibrated under the null when 80% of the sample is used for fitting, but the remaining small testing sample leads to very low power. In Appendix K we further compare the performance of our tests with additional baselines, constructed from classical variance and quantile-regression diagnostics.

5 Discussion

We propose a novel Hilbert-valued debiasing approach for inference and testing on kernel measures of residual noise heterogeneity, motivated by the classical bivariate causal discovery problem studied in Peters et al. [2014], Mooij et al. [2016]. We use a Hilbert-valued efficient influence function to debias the kernel cross-covariance operator between covariates and nonparametrically estimated residuals, which allows us to overcome the difficulties posed by the fact that the scalar-valued test statistic has vanishing gradient at the null of independence. By establishing a Hilbert-valued central limit theorem with asymptotically minimal variance, we simultaneously establish consistency of the bootstrap estimate of the critical value, leading to an asymptotically valid test, as well as obtain a confidence interval for the scalar-valued test statistic of asymptotically optimal width under the alternative. We further use the reverse triangular inequality to derive a confidence interval which is everywhere valid. We validate the performance of our novel inferential procedure in a range of synthetic experiments, demonstrating convincing gains over baselines which employ sample splitting between nuisance fitting and inference and the plug-in principle, where our method consistently achieves Type I error close to the nominal level and is powerful against alternatives.

A number of areas warrant further investigation. Firstly, it would be valuable to develop a non-asymptotic theory for our estimators and confidence sets. A second direction concerns the choice of centering. Using U-statistics improves coverage in some experiments, but this advantage over asymptotically equivalent V-statistics is not uniform. The use of U-statistic estimators of $\|\Psi_0\|^2$ does not trivially extend to the more conservative confidence sets based on the reverse triangular inequality, but it would be interesting to understand how their finite-sample performance can be optimized.

While we use the squared RKHS norm to map the Hilbert-valued one-step estimator to a scalar-valued object, we could in principle do this via any quadratic form. A particularly natural choice is to use the regularized inverse of the infinite-dimensional covariance operator, as suggested by Luedtke and Chung [2024], which takes into account which directions are easier to estimate from the data. Further investigation of this matter could yield RTI-style confidence sets with better coverage.

Disclosure of Funding

J.W., D.M. and A.G. are supported by the Gatsby Charitable Foundation. Z.S. is supported by EPSRC, grant EP/Y022300/1. The authors declare no competing interests.

References

M. Albert, B. Laurent, A. Marrel, and A. Meynaoui. Adaptive test of independence based on HSIC measures. ***The Annals of Statistics***, 50(2):858 – 879, 2022.

J.-P. Aubin. ***Applied functional analysis***. John Wiley & Sons, 2000.

A. Berlinet and C. Thomas-Agnan. ***Reproducing Kernel Hilbert Spaces in Probability and Statistics***. Springer, 2011.

T. B. Berrett and R. J. Samworth. Nonparametric independence testing via mutual information. ***Biometrika***, 106(3):547–566, 2019.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. ***Efficient and Adaptive Estimation for Semiparametric Models***. Springer, New York, NY, 1998. ISBN 978-0-387-98473-5.

G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. ***Foundations of Computational Mathematics***, 18(4):971–1013, 2018. 0.1007/s10208-017-9359-7. ttps://doi.org/10.1007/s10208-017-9359-7.

T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. ***Econometrica***, 47(5):1287–1294, 1979. 0.2307/1911963.

M. B. Brown and A. B. Forsythe. Robust tests for the equality of variances. ***Journal of the American Statistical Association***, 69(346):364–367, 1974. 0.1080/01621459.1974.10482955.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. ***Foundations of Computational Mathematics***, 7(3):331–368, 2007.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. ***Analysis and Applications***, 4(04):377–408, 2006.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. ***The Econometrics Journal***, 21(1), 2018.

K. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, ***Advances in Neural Information Processing Systems***, volume 27. Curran Associates, Inc., 2014.

B. Efron. Bootstrap methods: another look at the jackknife. In ***Breakthroughs in statistics: Methodology and distribution***, pages 569–593. Springer, 1992.

J. Fawkes, R. Hu, R. J. Evans, and D. Sejdinovic. Doubly robust kernel statistics for testing distributional treatment effects. ***Transactions on Machine Learning Research***, 2024. ISSN 2835-8856.

S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. ***Journal of Machine Learning Research***, 21(205):1–38, 2020. ttp://jmlr.org/papers/v21/19-734.html.

A. Gretton. A simpler condition for consistency of a kernel independence test. ***arXiv preprint arXiv:1501.06103***, 2015.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, ***Algorithmic Learning Theory***, pages 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31696-1.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, ***Advances in Neural Information Processing Systems***, volume 20. Curran Associates, Inc., 2007.

Z. He, R. Pogodin, Y. Li, N. Deka, A. Gretton, and D. J. Sutherland. On the hardness of conditional independence testing in practice. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, ***Advances in Neural Information Processing Systems***, volume 38, Main Conference, pages 172300–172339. Curran Associates, Inc., 2025. 0.52202/085713-5733.

R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of association based on ranks of distances. ***Biometrika***, 100(2):503–510, 2013.

O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt. Demystifying statistical learning based on efficient influence functions. ***The American Statistician***, 76(3):292–304, 2022.

Z. Huang and Z. Niu. Computationally efficient goodness-of-fit tests through kernelized stein discrepancy, 2026. ttps://arxiv.org/abs/2512.20007.

W. Jitkrittum, Z. Szabó, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. In D. Precup and Y. W. Teh, editors, ***Proceedings of the 34th International Conference on Machine Learning***, volume 70 of ***Proceedings of Machine Learning Research***, pages 1742–1751. PMLR, 2017.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. In ***Handbook of statistical methods for precision medicine***, pages 207–236. Chapman and Hall/CRC, 2024.

O. Key, A. Gretton, F.-X. Briol, and T. Fernandez. Composite goodness-of-fit tests with kernels. ***Journal of Machine Learning Research***, 26(51):1–60, 2025. ttp://jmlr.org/papers/v26/24-0276.html.

I. Kim and A. Ramdas. Dimension-agnostic inference using cross U-statistics. ***Bernoulli***, 30(1):683 – 711, 2024. 0.3150/23-BEJ1613. ttps://doi.org/10.3150/23-BEJ1613.

R. Koenker. A note on studentizing a test for heteroscedasticity. ***Journal of Econometrics***, 17(1):107–112, 1981. ISSN 0304-4076. ttps://doi.org/10.1016/0304-4076(81)90062-2. ttps://www.sciencedirect.com/science/article/pii/0304407681900622.

R. Koenker and G. Bassett, Jr. Regression quantiles. ***Econometrica***, 46(1):33–50, 1978. 0.2307/1913643.

A. J. Lee. ***U-Statistics: Theory and Practice***. Routledge, 1 edition, 1990. 0.1201/9780203734520. ttps://doi.org/10.1201/9780203734520.

Y. Li, R. Pogodin, D. J. Sutherland, and A. Gretton. Self-supervised learning with kernel dependence maximization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, ***Advances in Neural Information Processing Systems***, volume 34, pages 15543–15556. Curran Associates, Inc., 2021.

Z. Li, D. Meunier, M. Mollenhauer, and A. Gretton. Towards optimal Sobolev norm rates for the vector-valued regularized least-squares algorithm. ***Journal of Machine Learning Research***, 25 (181):1–51, 2024. ttp://jmlr.org/papers/v25/23-1663.html.

J. Lin and V. Cevher. Optimal distributed learning with multi-pass stochastic gradient methods. In J. Dy and A. Krause, editors, ***Proceedings of the 35th International Conference on Machine Learning***, volume 80 of ***Proceedings of Machine Learning Research***, pages 3092–3101. PMLR, 10–15 Jul 2018.

J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. ***Applied and Computational Harmonic Analysis***, 48(3):868–890, 2020.

A. Luedtke. Simplifying debiased inference via automatic differentiation and probabilistic programming. ***Journal of the Royal Statistical Society Series B: Statistical Methodology***, 88(1):313–329, 02 2026. ISSN 1369-7412. 0.1093/jrsssb/qkaf052.

A. Luedtke and I. Chung. One-step estimation of differentiable Hilbert-valued parameters. ***The Annals of Statistics***, 52(4):1534 – 1563, 2024. 0.1214/24-AOS2403. ttps://doi.org/10.1214/24-AOS2403.

D. Martinez-Taboada, A. Ramdas, and E. Kennedy. An efficient doubly-robust test for the kernel treatment effect. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, ***Advances in Neural Information Processing Systems***, volume 36, pages 59924–59952. Curran Associates, Inc., 2023. 0.52202/075280-2619.

A. McClean, S. Balakrishnan, E. H. Kennedy, and L. Wasserman. Double cross-fit doubly robust estimators: Beyond series regression. ***Journal of the Royal Statistical Society Series B: Statistical Methodology***, 88(4):1469–1491, 09 2026. ISSN 1369-7412. 0.1093/jrsssb/qkag057. ttps://doi.org/10.1093/jrsssb/qkag057.

S. McGrath and R. Mukherjee. Nuisance function tuning and sample splitting for optimally estimating a doubly robust functional, 2026. ttps://arxiv.org/abs/2212.14857.

S. Minsker. On some extensions of bernstein's inequality for self-adjoint operators. ***Statistics & Probability Letters***, 127:111–119, 2017. ISSN 0167-7152.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. ***Journal of Machine Learning Research***, 17 (32):1–102, 2016.

P. Morzywolek, P. B. Gilbert, and A. Luedtke. Inference on local variable importance measures for heterogeneous treatment effects. ***arXiv preprint arXiv:2510.18843***, 2025.

J. Park and K. Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, ***Advances in Neural Information Processing Systems***, volume 33, pages 21247–21259. Curran Associates, Inc., 2020.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. ***Journal of Machine Learning Research***, 15(58):2009–2053, 2014.

R. Pogodin, A. Schrab, Y. Li, D. J. Sutherland, and A. Gretton. Practical kernel tests of conditional independence. ***arXiv preprint arXiv:2402.13196***, 2024.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, ***Advances in Neural Information Processing Systems***, volume 20. Curran Associates, Inc., 2007.

A. Rotnitzky, E. Smucler, and J. M. Robins. Characterization of parameters with a mixed bias property. ***Biometrika***, 108(1):231–238, 03 2021. ISSN 0006-3444. 0.1093/biomet/asaa054. ttps://doi.org/10.1093/biomet/asaa054.

A. Schick. On Asymptotically Efficient Estimation in Semiparametric Models. ***The Annals of Statistics***, 14(3):1139 – 1151, 1986.

A. Schrab, I. Kim, B. Guedj, and A. Gretton. Efficient aggregated kernel tests using incomplete u-statistics. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, ***Advances in Neural Information Processing Systems***, volume 35, pages 18793–18807. Curran Associates, Inc., 2022. 0.52202/068431-1365.

R. J. Serfling. ***Approximation theorems of mathematical statistics***. John Wiley & Sons, 2009.

S. Shekhar, I. Kim, and A. Ramdas. A permutation-free kernel independence test. ***Journal of Machine Learning Research***, 24(369):1–68, 2023. ttp://jmlr.org/papers/v24/23-0248.html.

H. Shen, S. Jegelka, and A. Gretton. Fast kernel-based independent component analysis. ***IEEE Transactions on Signal Processing***, 57:3498–3511, 2009.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In V. Corruble, M. Takeda, and E. Suzuki, editors, ***Discovery Science***, pages 40–41, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75488-6.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. ***Journal of Machine Learning Research***, 13(47):1393–1434, 2012. ttp://jmlr.org/papers/v13/song12a.html.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. ***Journal of Machine Learning Research***, 11 (50):1517–1561, 2010. ttp://jmlr.org/papers/v11/sriperumbudur10a.html.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. ***Journal of Machine Learning Research***, 12(70):2389–2410, 2011. ttp://jmlr.org/papers/v12/sriperumbudur11a.html.

I. Steinwart and A. Christmann. ***Support vector machines***. Springer Science & Business Media, 2008.

A. B. Tsybakov. ***Introduction to nonparametric estimation***. Springer, 2009.

A. W. van der Vaart. ***Asymptotic statistics***, volume 3 of ***Cambridge Series in Statistical and Probabilistic Mathematics***. Cambridge university press, 2000.

W. Wang, M. Gong, B. Huang, J. Bailey, B. Han, K. Zhang, and F. Liu. Practical kernel selection for kernel-based conditional independence test. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, ***Advances in Neural Information Processing Systems***, volume 38, Main Conference, pages 49333–49370. Curran Associates, Inc., 2025. 0.52202/085713-1650.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. ***Econometrica***, 48(4):817–838, 1980. 0.2307/1912934.

C. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In T. Leen, T. Dietterich, and V. Tresp, editors, ***Advances in Neural Information Processing Systems***, volume 13. MIT Press, 2000.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. ***Statistics and Computing***, 28(1):113–130, 2018.

A Guide to appendix

In this section, we provide a guide to the contents of the appendix.

Sample splitting allows us to avoid restrictive Donsker-type complexity restrictions on function classes, and significantly simplify proofs. We refer the reader to Kennedy [2024, Section 4.2] for an exposition on sample splitting and cross-fitting. The main drawback of the sample splitting estimator is that it only uses half of the total samples for evaluation, therefore its asymptotic variance is two times larger than the semiparametric efficiency bound for the full sample. The virtue of the sample splitting estimator is that it is significantly less notationally burdensome to present, as there is no need to perform bookkeeping for the fold indices. In order to strike a balance between pedagogical clarity and generality, we opted to present the sample splitting estimator, as it already reveals the essential features of our debiased procedure and the nuisance requirements (Requirements 1B and 1C).

As explained in Remark 1, by swapping the role of evaluation and fitting samples, we can recover full statistical efficiency by using a K-fold cross-fitted estimator. The prototypical example is the 2-fold cross-fitted estimator. It is standard in theoretical works on semiparametric efficiency theory to present guarantees for the 2-fold cross-fitted estimator, as the extension of the proofs to the K-fold cross-fitted one presents no essential difficulty while requiring more burdensome notation, see e.g. Luedtke and Chung [2024, Section 2.5]. Furthermore, the asymptotic variance of the 2-fold cross-fitted estimator already attains the semiparametric efficiency bound.

The contents of the appendix can therefore be summarized as follows. In Appendix B, we introduce additional background material. In Appendix C we prove Theorem 1 in the main text, in the more general setting where Assumption 1 may not hold. Our results therein reduce immediately to Theorem 1 when Assumption 1 does in fact hold.

In Appendix D, we generalize the results in Luedtke and Chung [2024] for generic Hilbert-valued parameters to a two-fold cross-fitted variant of the estimator presented in Eq. (2).

The K-fold cross-fitted estimator often enjoys superior finite sample performance to the 2-fold cross-fitted estimator, and as a result, in all of our empirical simulations we use a K-fold cross-fitted estimator for $K \geq 2$. To support this and to ensure reproducibility of our research, we derive the closed form solution for the K-fold cross-fitted estimator in Appendix F, where we are fully explicit about how to perform tensor slice operations that let us perform cross-fitting in full generality.

In Appendix G, we provide supplementary theoretical results to the two methods of constructing confidence intervals proposed in Section 3.2 in the main text, for the general K-fold cross-fitting setting. In particular, we provide a closed form estimator $\widehat{\sigma}^2$ for the asymptotic variance in the delta method in the K-fold cross-fitting case, fully reusing expressions computed in Appendix F. We then establish the consistency for $\widehat{\sigma}^2$. In Appendix H, we construct the data-adaptive diagnostic for the trustworthiness of the delta method introduced in Section 3.2. In Appendix E, we provide a detailed introduction into statistical learning theory for vector-valued kernel ridge regression, which allows us to obtain convergence rates for the nuisance $\widehat{v}_j$, and thereby show when Requirements 1B and 1C are satisfied by more fundamental assumptions on Sobolev smoothness.

In Appendix J, we provide all experimental results not included in the main text and experimental configuration. Appendix I and Appendix L contain material not essential for exposition of our main method, but may help the reader contextualize our paper in the broader kernel testing literature. They respectively address comparison with cross U-statistic [Kim and Ramdas, 2024] and interpretable kernel testing via witness function learning.

B Notation and background

Notation	Meaning
$O = (X, W, Y), o_i = (x_i, w_i, y_i)$	Observation and realized observation.
$\mathcal{M}$, P_0, P	Nonparametric model, true law, and generic law. Except for H_0, subscript 0 always denotes an object indexed by the true data generating process.
m_P, ξ_P	Regression function $m_P(w) = \mathbb{E}_P[Y \mid W = w]$ and $\xi_P(w, y) = y - m_P(w)$.
k, l; $\mathcal{H}_k, \mathcal{H}_l, \mathcal{H}_{kl}$	Reproducing kernels on $\mathcal{X}$ and $\mathbb{R}^{d_y}$; corresponding reproducing kernel Hilbert spaces; tensor-product RKHS $\mathcal{H}_k \otimes \mathcal{H}_l$.
φ_k, φ_l, $\varphi_{\xi,P}$	Canonical feature maps, with $\varphi_{\xi,P}(w, y) = \varphi_l\{\xi_P(w, y)\}$.
$\mu_{X,P}$, $\mu_{\xi,P}$	Kernel mean embeddings $\mathbb{E}_P[\varphi_k(X)]$ and $\mathbb{E}_P[\varphi_{\xi,P}(W, Y)]$.
$\widetilde{\varphi}_{k,P}$, $\widetilde{\varphi}_{\xi,P}$	Centered feature maps obtained by subtracting the corresponding kernel mean embeddings.
$\Psi(P)$, Ψ_0	RKHS cross-covariance operator between X and the residual $\xi_P(W, Y)$; $\Psi_0 = \Psi(P_0)$.
$\partial_j \varphi_l$	Derivative of the residual feature map with respect to residual coordinate j.
$\widehat{\cdot}$, $\check{\cdot}$, $\cdot^{(-k)}$	Estimated object, cross-fitted object, and object fitted without fold k.

This table records only notation used across the paper. Section-specific notation is introduced again, locally, in the appendix section where it is used.

B.1 Background on semiparametric efficiency theory

Pathwise differentiability for a Hilbert-valued parameter. Let $\mathcal{P}$ be a collection of distributions defined on a common complete metric space $(\mathcal{Z}, \mathcal{B})$. Let $\mathcal{H}$ denote a reproducing kernel Hilbert space over a set $\mathcal{T}$ and $\nu : \mathcal{P} \to \mathcal{H}$ a parameter whose value is to be estimated. We let the feature map of $\mathcal{H}$ be denoted as ϕ. The parameter ν is said to be pathwise differentiable at $P \in \mathcal{P}$ if and only if there exists a continuous linear operator $\dot{\nu}_P : \dot{\mathcal{P}}_P \to \mathcal{H}$ such that for all parametric submodels $\{P_\epsilon : \epsilon \in [0, \delta)\}$ with score function s and centred at P with $P_0 = P$, we have

$$\|\nu(P_\epsilon) - \nu(P) - \epsilon \dot{\nu}_P(s)\|_{\mathcal{H}} = o(\epsilon).$$

The operator $\dot{\nu}_P$ is called the local parameter of ν at P and its Hermitian adjoint $\dot{\nu}_P^*$ is called the efficient influence operator. We refer the reader to Luedtke and Chung [2024, Section 2.2] for a full definition of a parametric submodel and a score function.

Hilbert-valued one-step estimator. In the notation of the above paragraph, we define $\phi_P : \mathcal{Z} \to \mathcal{H}$ as follows for each $t \in \mathcal{T}$:

$$\phi_P(z)(t) = \dot{\nu}_P^*(\phi(t))(z), \qquad P\text{-a.s. } z.$$

If ϕ_P is square integrable in the sense of $\|\phi_P\|_{L^2(P;\mathcal{H})} < \infty$, then we say that ϕ_P is the EIF of ν. In this case, we can then construct a one-step estimator of the form $\nu(\widehat{P}) + P_n \phi_{\widehat{P}}$, where P_n denotes the sample average and $\widehat{P}$ is an initial data-driven estimate of the data-generating distribution P_0.

B.2 Background on reproducing kernel Hilbert spaces

Lemma 1. *Let $l : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ be a symmetric positive definite kernel with canonical feature map $\varphi_l : \mathbb{R}^{d_y} \to \mathcal{H}_l$. Suppose that $\sup_u l(u, u) \leq \kappa_l^2$ for a finite constant $\kappa_l \geq 0$, that φ_l is continuously*

Fréchet differentiable, and that

$$\|D\varphi_l(u) - D\varphi_l(v)\|_{\mathrm{op}} \le L_2\|u - v\|, \qquad u, v \in \mathbb{R}^{d_y},$$

for a finite constant $L_2 \ge 0$. Then, for all $u, v, h \in \mathbb{R}^{d_y}$,

$$\sup_u \|D\varphi_l(u)\|_{\mathrm{op}} \le \sqrt{2\kappa_l L_2},$$
$$\|\varphi_l(u) - \varphi_l(v)\|_{\mathcal{H}_l} \le \sqrt{2\kappa_l L_2}\,\|u - v\|,$$
$$\|\varphi_l(u+h) - \varphi_l(u) - D\varphi_l(u)[h]\|_{\mathcal{H}_l} \le \frac{L_2}{2}\|h\|^2.$$

For every $j, r \in \{1, \dots, d_y\}$, the mixed derivative $\partial_{1,j}\partial_{2,r} l$ exists, is continuous, and satisfies

$$\langle \partial_j \varphi_l(u), \partial_r \varphi_l(v)\rangle_{\mathcal{H}_l} = \partial_{1,j}\partial_{2,r} l(u, v),$$

where $\partial_{1,j}$ and $\partial_{2,r}$ differentiate the first and second arguments of l, respectively. For every $a \in \mathbb{R}^{d_y}$, the derivative satisfies

$$D\varphi_l(u)[a] = \sum_{j=1}^{d_y} a_j \partial_j \varphi_l(u). \tag{7}$$

Proof. The reproducing identity gives $\|\varphi_l(u)\|^2_{\mathcal{H}_l} = l(u, u) \le \kappa_l^2$. By the Banach-valued fundamental theorem of calculus and the Lipschitz condition on $D\varphi_l$,

$$\begin{aligned}
&\|\varphi_l(u+h) - \varphi_l(u) - D\varphi_l(u)[h]\|_{\mathcal{H}_l} \\
&\quad = \left\| \int_0^1 \{D\varphi_l(u+th) - D\varphi_l(u)\}[h]\, dt \right\|_{\mathcal{H}_l} \\
&\quad \le \int_0^1 L_2 t\|h\|^2\, dt = \frac{L_2}{2}\|h\|^2.
\end{aligned}$$

For a unit vector $a \in \mathbb{R}^{d_y}$ and $t > 0$, applying this bound to ta and $-ta$ and using the triangle inequality yields

$$\begin{aligned}
\|D\varphi_l(u)[a]\|_{\mathcal{H}_l} &\le \frac{\|\varphi_l(u+ta) - \varphi_l(u-ta)\|_{\mathcal{H}_l}}{2t} + \frac{L_2 t}{2} \\
&\le \frac{\kappa_l}{t} + \frac{L_2 t}{2}.
\end{aligned}$$

Taking the infimum over $t > 0$ gives

$$\sup_u \|D\varphi_l(u)\|_{\mathrm{op}} \le \inf_{t>0}\left(\frac{\kappa_l}{t} + \frac{L_2 t}{2}\right) = \sqrt{2\kappa_l L_2}.$$

The Lipschitz bound on φ_l follows by integrating $D\varphi_l$ along the segment from u to v. Differentiating

$$l(u, v) = \langle \varphi_l(u), \varphi_l(v)\rangle_{\mathcal{H}_l}$$

once in each argument gives the mixed derivative identity and the continuity of $\partial_{1,j}\partial_{2,r} l$. The formula in Eq. (7) for $D\varphi_l(u)$ follows from its linearity. □

Lemma 2 (Sufficient conditions on derivatives of the kernel)**.** *Let $l : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ be a symmetric positive definite kernel satisfying $\sup_u l(u, u) \le \kappa_l^2$ for a finite constant $\kappa_l \ge 0$. For multi-indices $\alpha, \beta \in \mathbb{N}_0^{d_y}$, let $\partial_1^\alpha \partial_2^\beta$ denote differentiation in the first and second arguments, respectively. Suppose that $\partial_1^\alpha \partial_2^\beta l$ exists and is continuous whenever $|\alpha| \le 2$ and $|\beta| \le 2$. For a direction $a \in \mathbb{R}^{d_y}$, let $\partial_{1,a}$ and $\partial_{2,a}$ denote the corresponding directional derivatives. Suppose that, for a finite constant $L_2 \ge 0$,*

$$\sup_{u \in \mathbb{R}^{d_y}} \sup_{\substack{a,b \in \mathbb{R}^{d_y} \\ \|a\| = \|b\| = 1}} \left(\partial_{1,a}\partial_{1,b}\partial_{2,a}\partial_{2,b} l\right)(u, u) \le L_2^2.$$

The derivatives in this bound are taken before restricting to the diagonal. Then the canonical feature map $\varphi_l : \mathbb{R}^{d_y} \to \mathcal{H}_l$ is twice continuously Fréchet differentiable and

$$\sup_u \|D^2\varphi_l(u)\|_{\mathrm{op}} \le L_2.$$

In particular, l satisfies Assumption 4.

Proof. Applying the differentiability theorem for reproducing kernel Hilbert spaces [Steinwart and Christmann, 2008, Lemma 4.34 and Corollary 4.36] iteratively gives the continuous second derivative of φ_l and, for all $u, a, b \in \mathbb{R}^{d_y}$,

$$\begin{aligned}\|D^2\varphi_l(u)[a,b]\|_{\mathcal{H}_l}^2 &= \big(\partial_{1,a}\partial_{1,b}\partial_{2,a}\partial_{2,b}l\big)(u,u)\\ &\le L_2^2\|a\|^2\|b\|^2.\end{aligned}$$

Taking the supremum over unit directions gives the bound on $D^2\varphi_l$. The Banach-valued fundamental theorem of calculus gives

$$\begin{aligned}\|D\varphi_l(u) - D\varphi_l(v)\|_{\mathrm{op}} &= \left\|\int_0^1 D^2\varphi_l\{v + t(u-v)\}[u-v,\cdot]\,dt\right\|_{\mathrm{op}}\\ &\le L_2\|u-v\|,\end{aligned}$$

which proves Assumption 4. □

Proposition 1 (Gaussian and Sobolev kernels)**.** *For every $\sigma > 0$, the Gaussian kernel on $\mathbb{R}^{d_y}$,*

$$l_\sigma(u,v) = \exp\{-\|u-v\|^2/(2\sigma^2)\},$$

has $l_\sigma(u,u) = 1$ and satisfies Assumption 4 with $L_2 = \sqrt{3}\,\sigma^{-2}$. Its canonical feature map also satisfies

$$\sup_u \|D\varphi_{l_\sigma}(u)\|_{\mathrm{op}} \le \sigma^{-1}.$$

For every $\rho > 0$ and $s > d_y/2 + 2$, the Sobolev kernel

$$l_{s,\rho}(u,v) = \frac{1}{(2\pi)^{d_y}}\int_{\mathbb{R}^{d_y}} e^{i\omega^\top(u-v)}(1+\rho^2\|\omega\|^2)^{-s}\,d\omega$$

is bounded and satisfies Assumption 4. Matérn kernels with smoothness parameter $\nu > 2$ and any positive scale satisfy the same conditions.

Proof. The Gaussian kernel is positive definite, is smooth, and has $l_\sigma(u,u) = 1$. Direct differentiation gives, for unit vectors $a, b \in \mathbb{R}^{d_y}$,

$$\big(\partial_{1,a}\partial_{2,a}l_\sigma\big)(u,u) = \sigma^{-2}$$

and

$$\big(\partial_{1,a}\partial_{1,b}\partial_{2,a}\partial_{2,b}l_\sigma\big)(u,u) = \sigma^{-4}\{1 + 2(a^\top b)^2\} \le 3\sigma^{-4}.$$

The second identity and Lemma 2 give the Lipschitz constant L_2. The first identity and Lemma 1 then give the bound on $D\varphi_{l_\sigma}$.

For the Sobolev kernel, positive definiteness follows from its nonnegative spectral density. Since $s > d_y/2 + 2$,

$$\int_{\mathbb{R}^{d_y}} (1 + \|\omega\|^4)(1+\rho^2\|\omega\|^2)^{-s}\,d\omega < \infty.$$

Dominated convergence gives continuous mixed derivatives through order two in each argument. For unit directions a, b, their diagonal values satisfy

$$\big(\partial_{1,a}\partial_{1,b}\partial_{2,a}\partial_{2,b}l_{s,\rho}\big)(u,u) \le \frac{1}{(2\pi)^{d_y}}\int_{\mathbb{R}^{d_y}} \|\omega\|^4(1+\rho^2\|\omega\|^2)^{-s}\,d\omega < \infty.$$

The zeroth spectral moment bounds $l_{s,\rho}(u,u)$, so Lemma 2 applies. The Matérn smoothness parameter is $\nu = s - d_y/2$, which gives the condition $\nu > 2$. □

C Efficient influence function of the cross-covariance operator

We derive the efficient influence function without assuming that X is a measurable function of W. Under Assumption 1, the resulting expression reduces to Theorem 1.

Proof of Theorem 1 in the general setting where we drop Assumption 1. Following a standard convention in semiparametric statistical theory, we use the subscript 0 to denote any objects indexed by the true distribution P_0. We have

$$\begin{aligned}
&\mathbb{E}_P\left[\widetilde{\varphi}_{k,P}(X) \otimes \widetilde{\varphi}_{\xi,P}(W,Y)\right] \\
&= \mathbb{E}_P[(\varphi_k(X) - \mathbb{E}_P[\varphi_k(X')]) \otimes (\varphi_{\xi,P}(W,Y) - \mathbb{E}_P[\varphi_{\xi,P}(W',Y')])] \\
&= \mathbb{E}_P[\varphi_k(X) \otimes \varphi_{\xi,P}(W,Y)] - \mathbb{E}_P[\varphi_k(X)] \otimes \mathbb{E}_P[\varphi_{\xi,P}(W,Y)].
\end{aligned}$$

We define the following RKHS-valued parameters on $\mathcal{M}$,

$$\begin{aligned}
\Psi_1(P) &:= \mathbb{E}_P[\varphi_k(X) \otimes \varphi_l(Y - \mathbb{E}_P[Y \mid W])] = \mathbb{E}_P[\varphi_k(X) \otimes \varphi_{\xi,P}(W,Y)] \\
\Psi_2(P) &:= \mathbb{E}_P[\varphi_k(X)] \\
\Psi_3(P) &:= \mathbb{E}_P[\varphi_l(Y - \mathbb{E}_P[Y \mid W])] = \mathbb{E}_P[\varphi_{\xi,P}(W,Y)] \\
\Psi(P) &= \Psi_1(P) - \Psi_2(P) \otimes \Psi_3(P).
\end{aligned}$$

We write $\Psi_1(P)$ using the computation graph of primitives

$$\begin{aligned}
\theta_1 : \mathcal{M} &\to L^2(P_0; \mathbb{R}^{d_y}) \\
P &\mapsto ((x,w,y) \mapsto y) \\
\theta_2 : \mathcal{M} \times L^2(P_0; \mathbb{R}^{d_y}) &\to L^2(P_{0,W}; \mathbb{R}^{d_y}) \\
(P, h_1) &\mapsto \mathbb{E}_P[h_1(X,W,Y) \mid W = w] \\
\theta_3 : \mathcal{M} \times L^2(P_{0,W}; \mathbb{R}^{d_y}) &\to L^2(P_0; \mathbb{R}^{d_y}) \\
(P, h_2) &\mapsto ((x,w,y) \mapsto h_2(w)) \\
\theta_4 : \mathcal{M} \times L^2(P_0; \mathbb{R}^{d_y}) &\to L^2(P_0; \mathbb{R}^{d_y}) \\
(P, h_3) &\mapsto ((x,w,y) \mapsto y - h_3(x,w,y)) \\
\theta_5 : \mathcal{M} \times L^2(P_0; \mathbb{R}^{d_y}) &\to L^2(P_0; \mathcal{H}_{kl}) \\
(P, h_4) &\mapsto ((x,w,y) \mapsto \varphi_k(x) \otimes \varphi_l(h_4(x,w,y))) \\
\theta_6 : \mathcal{M} \times L^2(P_0; \mathcal{H}_{kl}) &\to \mathcal{H}_{kl} \\
(P, h_5) &\mapsto \mathbb{E}_P[h_5(X,W,Y)].
\end{aligned}$$

We let $h_1 = \theta_1(P)$, $h_2 = \theta_2(P, h_1)$, $h_3 = \theta_3(P, h_2)$, $h_4 = \theta_4(P, h_3)$, $h_5 = \theta_5(P, h_4)$, $h_6 = \theta_6(P, h_5)$. Concretely, we have

$$\begin{aligned}
h_1(x,w,y) &= y \\
h_2(w) &= \mathbb{E}_P[Y \mid W = w] \\
h_3(x,w,y) &= h_2(w) = \mathbb{E}_P[Y \mid W = w] \\
h_4(x,w,y) &= y - \mathbb{E}_P[Y \mid W = w] \\
h_5(x,w,y) &= \varphi_k(x) \otimes \varphi_l(y - \mathbb{E}_P[Y \mid W = w]) \\
h_6 &= \mathbb{E}_P\left[\varphi_k(X) \otimes \varphi_l(Y - \mathbb{E}_P[Y \mid W])\right].
\end{aligned}$$

We restrict the inputs of θ_2 to functions bounded in norm by $1 + \|y\|$, and those of θ_6 to functions bounded in norm by $(1 + \kappa_k)\kappa_l$. Under Assumption 3, these envelopes satisfy the conditional moment condition in Luedtke [2026, Appendix C.2.2, Lemma S8], with reference law P_0. We have $\Psi_1(P) = h_6$. Note that all of the above primitives are contained in Luedtke [2026, Table 1]. They are respectively ***constant map***, ***conditional mean***, ***lift to new domain***, ***pointwise operation***, ***pointwise operation*** and ***marginal mean***. For $\Psi_1(P)$, we initialize the adjoint variables $f_i = 0$ for $0 \le i \le 5$ and choose $f_6 \in \mathcal{H}_{kl}$. We use the notation convention in Luedtke and Chung [2024] to use subscript 0 to denote objects associated with the true data generating process. We let $\dot{\mathcal{M}}_0$ denote the tangent space of the statistical model $\mathcal{M}$ at P_0, which we assume to be the space $L_0^2(P_0)$ of mean-zero square integrable functions. This corresponds to a nonparametric statistical model. In the following calculation, we evaluate all intermediate functions $h_1, \ldots, h_6$ at P_0. We now describe the backpropagation procedure.

1. $j = 6$. h_6 is a ***marginal mean***. We have that the pathwise derivative of θ_6 at (P_0, h_5) is given by

$$D[\theta_6](P_0, h_5) : \dot{\mathcal{M}}_0 \oplus L^2(P_0; \mathcal{H}_{kl}) \to \mathcal{H}_{kl}.$$

We backpropagate f_6 through θ_6 to f_0, f_5 as follows:

$$\begin{aligned}\begin{pmatrix} f_0 \\ f_5 \end{pmatrix} &+= D[\theta_6](P_0, h_5)^*(f_6) \\ &= \begin{pmatrix} (x, w, y) \mapsto \langle f_6, h_5(x, w, y) - \mathbb{E}_0[h_5(X, W, Y)] \rangle_{\mathcal{H}_{kl}} \\ (x, w, y) \mapsto f_6 \end{pmatrix}.\end{aligned}$$

2. $j = 5$. We have

$$D[\theta_5](P_0, h_4) : \dot{\mathcal{M}}_0 \oplus L^2(P_0; \mathbb{R}^{d_y}) \to L^2(P_0; \mathcal{H}_{kl}).$$

We backpropagate f_5 through θ_5 to f_4 as follows:

$$\begin{aligned}\begin{pmatrix} f_0 \\ f_4 \end{pmatrix} &+= D[\theta_5](P_0, h_4)^*(f_5) \\ &= \begin{pmatrix} 0 \\ (x, w, y) \mapsto (\langle f_5(x, w, y), \varphi_k(x) \otimes (\partial_j \varphi_l)(h_4(x, w, y)) \rangle_{\mathcal{H}_{kl}})_{j=1}^{d_y} \end{pmatrix},\end{aligned}$$

where $\partial_j \varphi_l$ refers to the jth-entry of the gradient of $\varphi_l : \mathbb{R}^{d_y} \to \mathcal{H}_l$.

3. $j = 4$. We have

$$D[\theta_4](P_0, h_3) : \dot{\mathcal{M}}_0 \oplus L^2(P_0; \mathbb{R}^{d_y}) \to L^2(P_0; \mathbb{R}^{d_y}).$$

We backpropagate f_4 through θ_4 to f_3 as follows:

$$\begin{pmatrix} f_0 \\ f_3 \end{pmatrix} += D[\theta_4](P_0, h_3)^*(f_4) = \begin{pmatrix} 0 \\ (x, w, y) \mapsto -f_4(x, w, y) \end{pmatrix}.$$

4. $j = 3$. We have

$$D[\theta_3](P_0, h_2) : \dot{\mathcal{M}}_0 \oplus L^2(P_{0,W}; \mathbb{R}^{d_y}) \to L^2(P_0; \mathbb{R}^{d_y}).$$

We backpropagate f_3 through θ_3 to f_2 as follows:

$$\begin{pmatrix} f_0 \\ f_2 \end{pmatrix} += D[\theta_3](P_0, h_2)^*(f_3) = \begin{pmatrix} 0 \\ w \mapsto \mathbb{E}_0[f_3(X, W, Y) \mid W = w] \end{pmatrix}.$$

5. $j = 2$. We have

$$D[\theta_2](P_0, h_1) : \dot{\mathcal{M}}_0 \oplus L^2(P_0; \mathbb{R}^{d_y}) \to L^2(P_{0,W}; \mathbb{R}^{d_y}).$$

We backpropagate f_2 through θ_2 to f_1 as follows:

$$\begin{aligned}\begin{pmatrix} f_0 \\ f_1 \end{pmatrix} &+= D[\theta_2](P_0, h_1)^*(f_2) \\ &= \begin{pmatrix} (x, w, y) \mapsto f_2(w)^\top (h_1(x, w, y) - \mathbb{E}_0[h_1(X, W, Y) \mid W = w]) \\ (x, w, y) \mapsto f_2(w) \end{pmatrix}.\end{aligned}$$

6. $j = 1$. We have

$$D[\theta_1](P_0) : \dot{\mathcal{M}}_0 \to L^2(P_0; \mathbb{R}^{d_y}).$$

We backpropagate f_1 through θ_1 to f_0 as follows:

$$f_0 += D[\theta_1](P_0)^*(f_1) = 0.$$

We have

$$
\begin{aligned}
f_5(x,w,y) &= f_6 \\
f_4(x,w,y) &= (\langle f_6, \varphi_k(x) \otimes (\partial_j \varphi_l)(h_4(x,w,y))\rangle_{\mathcal{H}_{kl}})_{j=1}^{d_y} \\
f_3(x,w,y) &= -(\langle f_6, \varphi_k(x) \otimes (\partial_j \varphi_l)(h_4(x,w,y))\rangle_{\mathcal{H}_{kl}})_{j=1}^{d_y} \\
f_2(w) &= -\Big(\mathbb{E}_0\Big[\langle f_6, \varphi_k(X) \otimes (\partial_j \varphi_l)(h_4(X,W,Y))\rangle_{\mathcal{H}_{kl}} \\
&\qquad\qquad \Big| \; W = w\Big]\Big)_{j=1}^{d_y}, \\
f_1(x,w,y) &= -\Big(\mathbb{E}_0\Big[\langle f_6, \varphi_k(X) \otimes (\partial_j \varphi_l)(h_4(X,W,Y))\rangle_{\mathcal{H}_{kl}} \\
&\qquad\qquad \Big| \; W = w\Big]\Big)_{j=1}^{d_y}.
\end{aligned}
$$

Finally, summing the adjoint contributions for f_0 yields

$$
\begin{aligned}
f_0(x,w,y) &= \langle f_6, h_5(x,w,y) - \mathbb{E}_0[h_5(X,W,Y)]\rangle_{\mathcal{H}_{kl}} \\
&\quad + f_2(w)^\top \{h_1(x,w,y) - \mathbb{E}_0[h_1(X,W,Y) \mid W = w]\} \\
&= \langle f_6, \varphi_k(x) \otimes \varphi_{\xi,0}(w,y)\rangle_{\mathcal{H}_{kl}} \\
&\quad - \langle f_6, \mathbb{E}_0[\varphi_k(X) \otimes \varphi_{\xi,0}(W,Y)]\rangle_{\mathcal{H}_{kl}} \\
&\quad - \sum_{j=1}^{d_y} \mathbb{E}_0\big[\langle f_6, \varphi_k(X) \otimes (\partial_j \varphi_l)(\xi_0(W,Y))\rangle_{\mathcal{H}_{kl}} \mid W = w\big] \xi_{0,j}(w,y).
\end{aligned}
$$

Recall we defined $m_0(w) = \mathbb{E}_0[Y \mid W = w]$, $\xi_0(w,y) = y - m_0(w)$, and $\varphi_{\xi,0}(w,y) = \varphi_l\{\xi_0(w,y)\}$. We write $\mu_{X,0} := \mathbb{E}_0[\varphi_k(X)]$ and $\mu_{\xi,0} := \mathbb{E}_0[\varphi_{\xi,0}(W,Y)]$. For each $j = 1, \ldots, d_y$, define

$$
\begin{aligned}
v_{0,j}(w) &:= \mathbb{E}_0[(\partial_j \varphi_l)(\xi_0(W,Y)) \mid W = w] \in \mathcal{H}_l, \\
a_j(w) &:= \mathbb{E}_0[\varphi_k(X) \otimes (\partial_j \varphi_l)(\xi_0(W,Y)) \mid W = w] \in \mathcal{H}_{kl}.
\end{aligned}
$$

By Luedtke [2026, Theorem 1], $\Psi_1(P)$ is pathwise differentiable at P_0, and its efficient influence operator $\dot{\Psi}_{1,0}^* : \mathcal{H}_{kl} \to \dot{\mathcal{M}}_0$ is given by

$$
\begin{aligned}
\dot{\Psi}_{1,0}^*(f)(x,w,y) &= \langle f, \varphi_k(x) \otimes \varphi_{\xi,0}(w,y) - \mathbb{E}_0[\varphi_k(X) \otimes \varphi_{\xi,0}(W,Y)]\rangle_{\mathcal{H}_{kl}} \\
&\quad - \sum_{j=1}^{d_y} \xi_{0,j}(w,y) \langle f, a_j(w)\rangle_{\mathcal{H}_{kl}} \\
&= \langle f, \Gamma_1(x,w,y)\rangle_{\mathcal{H}_{kl}},
\end{aligned}
$$

where we define

$$
\Gamma_1(x,w,y) := \varphi_k(x) \otimes \varphi_{\xi,0}(w,y) - \mathbb{E}_0[\varphi_k(X) \otimes \varphi_{\xi,0}(W,Y)] - \sum_{j=1}^{d_y} \xi_{0,j}(w,y) a_j(w).
$$

Note $\Psi_2(P)$ is a ***kernel mean embedding*** (see Luedtke [2026, C.2.5]), since

$$
\Psi_2(P) = \int_{\mathcal{X}} k(x,\cdot) P_X(dx).
$$

By Luedtke [2026, Lemma S11], $\Psi_2(P)$ is pathwise differentiable at P_0, and its efficient influence operator $\dot{\Psi}_{2,0}^* : \mathcal{H}_k \to \dot{\mathcal{M}}_0$ is given by

$$
\dot{\Psi}_{2,0}^*(f)(x,w,y) = f(x) - \mathbb{E}_0[f(X)] = \langle f, \Gamma_2(x)\rangle_{\mathcal{H}_k},
$$

where we define

$$
\Gamma_2(x) := \varphi_k(x) - \mu_{X,0}.
$$

Note that $\Psi_3(P)$ can be obtained by modifying the primitive θ_5 of $\Psi_1(P)$ to

$$\begin{aligned}\widetilde{\theta}_5 : \mathcal{M} \times L^2(P_0; \mathbb{R}^{d_y}) &\to L^2(P_0; \mathcal{H}_l) \\ (P, h_4) &\mapsto ((x, w, y) \mapsto \varphi_l(h_4(x, w, y))).\end{aligned}$$

We conclude that $\Psi_3(P)$ is pathwise differentiable at P_0, with efficient influence operator $\dot{\Psi}^*_{3,0} : \mathcal{H}_l \to \dot{\mathcal{M}}_0$ given by

$$\begin{aligned}\dot{\Psi}^*_{3,0}(f)(x, w, y) &= \langle f, \varphi_{\xi,0}(w, y) - \mu_{\xi,0} \rangle_{\mathcal{H}_l} - \sum_{j=1}^{d_y} \xi_{0,j}(w, y) \langle f, v_{0,j}(w) \rangle_{\mathcal{H}_l} \\ &= \langle f, \Gamma_3(x, w, y) \rangle_{\mathcal{H}_l},\end{aligned}$$

where we define

$$\Gamma_3(x, w, y) := \varphi_{\xi,0}(w, y) - \mu_{\xi,0} - \sum_{j=1}^{d_y} \xi_{0,j}(w, y) v_{0,j}(w).$$

By product rule, for $s \in L^2_0(P_0)$, we have

$$\dot{\Psi}_0[s] = \dot{\Psi}_{1,0}[s] - \dot{\Psi}_{2,0}[s] \otimes \mu_{\xi,0} - \mu_{X,0} \otimes \dot{\Psi}_{3,0}[s] \in \mathcal{H}_{kl}.$$

Let $f \in \mathcal{H}_{kl}$ be arbitrary. We have

$$\begin{aligned}\langle f, \dot{\Psi}_0[s] \rangle_{\mathcal{H}_{kl}} &= \langle \dot{\Psi}^*_0 f, s \rangle_{L^2(P_0)} \\ &= \langle f, \dot{\Psi}_{1,0}[s] - \dot{\Psi}_{2,0}[s] \otimes \mu_{\xi,0} - \mu_{X,0} \otimes \dot{\Psi}_{3,0}[s] \rangle_{\mathcal{H}_{kl}}.\end{aligned}$$

We have, for all elementary tensors $u \otimes v \in \mathcal{H}_{kl}$,

$$\begin{aligned}\langle u \otimes v, \dot{\Psi}_{2,0}[s] \otimes \mu_{\xi,0} \rangle_{\mathcal{H}_{kl}} &= \langle u, \dot{\Psi}_{2,0}[s] \rangle_{\mathcal{H}_k} \langle v, \mu_{\xi,0} \rangle_{\mathcal{H}_l} \\ &= \langle \dot{\Psi}^*_{2,0} u, s \rangle_{L^2(P_0)} \langle v, \mu_{\xi,0} \rangle_{\mathcal{H}_l} \\ &= \langle \dot{\Psi}^*_{2,0} (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu^*_{\xi,0})(u \otimes v), s \rangle_{L^2(P_0)}.\end{aligned}$$

Hence, extending by linearity and continuity, we have for all $f \in \mathcal{H}_{kl}$,

$$\begin{aligned}\langle f, \dot{\Psi}_{2,0}[s] \otimes \mu_{\xi,0} \rangle_{\mathcal{H}_{kl}} &= \langle \dot{\Psi}^*_{2,0} (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu^*_{\xi,0}) f, s \rangle_{L^2(P_0)} \\ \langle f, \mu_{X,0} \otimes \dot{\Psi}_{3,0}[s] \rangle_{\mathcal{H}_{kl}} &= \langle \dot{\Psi}^*_{3,0} (\mu^*_{X,0} \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l}) f, s \rangle_{L^2(P_0)}.\end{aligned}$$

Therefore,

$$\begin{aligned}\dot{\Psi}^*_0(f)(x, w, y) &= \dot{\Psi}^*_{1,0}(f)(x, w, y) \\ &\quad - \dot{\Psi}^*_{2,0}\big((\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu^*_{\xi,0}) f\big)(x, w, y) \\ &\quad - \dot{\Psi}^*_{3,0}\big((\mu^*_{X,0} \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l}) f\big)(x, w, y) \\ &= \langle f, \Gamma_1(x, w, y) \rangle_{\mathcal{H}_{kl}} \\ &\quad - \langle (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu^*_{\xi,0}) f, \Gamma_2(x) \rangle_{\mathcal{H}_k} \\ &\quad - \langle (\mu^*_{X,0} \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l}) f, \Gamma_3(x, w, y) \rangle_{\mathcal{H}_l} \\ &= \left\langle f, \begin{array}{c} \Gamma_1(x, w, y) - (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu^*_{\xi,0})^* \Gamma_2(x) \\ - (\mu^*_{X,0} \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l})^* \Gamma_3(x, w, y) \end{array} \right\rangle_{\mathcal{H}_{kl}}.\end{aligned}$$

By Assumption 2 and lemma 1 and conditional Jensen's inequality, for every $j = 1, \ldots, d_y$,

$$\|a_j(W)\|_{\mathcal{H}_{kl}} \le \kappa_k \sqrt{2\kappa_l L_2}, \qquad \|v_{0,j}(W)\|_{\mathcal{H}_l} \le \sqrt{2\kappa_l L_2}.$$

Since $\mathbb{E}_0 \|\xi_0(W, Y)\|^2 \le \mathbb{E}_0 \|Y\|^2 < \infty$, the expressions for Γ_1, Γ_2, and Γ_3 are square integrable. Their means are zero by $\mathbb{E}_0[\xi_0(W, Y) \mid W] = 0$. Tensoring with a fixed kernel mean embedding

is a bounded linear map, so the expression inside the final inner product belongs to $L_0^2(P_0; \mathcal{H}_{kl})$. Therefore, the parameter Ψ has an EIF at P_0 in the sense of Luedtke and Chung [2024], given by

$$\phi_0(x, w, y) = \Gamma_1(x, w, y) - (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*)^* \Gamma_2(x) - (\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l})^* \Gamma_3(x, w, y).$$

We compute

$$\begin{aligned}(\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*)^* \Gamma_2(x) &= (\mathrm{Id}_{\mathcal{H}_k} \otimes_{\mathrm{op}} \mu_{\xi,0}^*)^* (\varphi_k(x) - \mu_{X,0}) \\ &= (\varphi_k(x) - \mu_{X,0}) \otimes \mu_{\xi,0} \\ &= \widetilde{\varphi}_{k,0}(x) \otimes \mu_{\xi,0},\end{aligned}$$

where the second line can be verified by a simple adjoint identity. Similarly, we have

$$\begin{aligned}(\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l})^* \Gamma_3(x, w, y) &= (\mu_{X,0}^* \otimes_{\mathrm{op}} \mathrm{Id}_{\mathcal{H}_l})^* \left(\varphi_{\xi,0}(w, y) - \mu_{\xi,0} - \sum_{j=1}^{d_y} \xi_{0,j}(w, y) v_{0,j}(w) \right) \\ &= \mu_{X,0} \otimes \left(\varphi_{\xi,0}(w, y) - \mu_{\xi,0} - \sum_{j=1}^{d_y} \xi_{0,j}(w, y) v_{0,j}(w) \right).\end{aligned}$$

Hence we have

$$\begin{aligned}\mathcal{H}_{kl} \ni \phi_0(x, w, y) = \varphi_k(x) \otimes \varphi_{\xi,0}(w, y) - \mathbb{E}_0[\varphi_k(X) \otimes \varphi_{\xi,0}(W, Y)] - \sum_{j=1}^{d_y} \xi_{0,j}(w, y) a_j(w) \\ - \widetilde{\varphi}_{k,0}(x) \otimes \mu_{\xi,0} - \mu_{X,0} \otimes \left(\varphi_{\xi,0}(w, y) - \mu_{\xi,0} - \sum_{j=1}^{d_y} \xi_{0,j}(w, y) v_{0,j}(w) \right).\end{aligned}$$

Define

$$\begin{aligned}F_j(w) &:= a_j(w) - \mu_{X,0} \otimes v_{0,j}(w) \\ &= \mathbb{E}_0[\{\varphi_k(X) - \mu_{X,0}\} \otimes (\partial_j \varphi_l)(\xi_0(W, Y)) \mid W = w] \in \mathcal{H}_{kl}.\end{aligned}$$

Since

$$\Psi_0 = \mathbb{E}_0[\varphi_k(X) \otimes \varphi_{\xi,0}(W, Y)] - \mu_{X,0} \otimes \mu_{\xi,0},$$

we obtain

$$\phi_0(x, w, y) = \widetilde{\varphi}_{k,0}(x) \otimes (\varphi_{\xi,0}(w, y) - \mu_{\xi,0}) - \sum_{j=1}^{d_y} \xi_{0,j}(w, y) F_j(w) - \Psi_0.$$

Under Assumption 1, $\widetilde{\varphi}_{k,0}(X)$ is measurable with respect to W. Taking it outside the conditional expectation gives

$$\begin{aligned}F_j(W) &= \widetilde{\varphi}_{k,0}(X) \otimes \mathbb{E}_0[(\partial_j \varphi_l)(\xi_0(W, Y)) \mid W] \\ &= \widetilde{\varphi}_{k,0}(X) \otimes v_{0,j}(W).\end{aligned}$$

Substituting this identity into the preceding expression yields

$$\begin{aligned}\phi_0(x, w, y) = \widetilde{\varphi}_{k,0}(x) \otimes \left[\widetilde{\varphi}_{l,0}(\xi_0(w, y)) - \sum_{j=1}^{d_y} \xi_{0,j}(w, y) v_{0,j}(w) \right] \\ - \Psi_0,\end{aligned}$$

which is the efficient influence function in Theorem 1. □

D Asymptotic linearity of the two-fold cross-fitted estimator

Throughout this section, we impose Assumption 1. Since $X = \pi(W)$, we represent an observation as $O = (W, Y) \in \mathcal{W} \times \mathbb{R}^{d_y}$. Fix $P \in \mathcal{M}$. Recall that

$$m_P(w) = \mathbb{E}_P[Y \mid W = w], \qquad \xi_P(w, y) = y - m_P(w),$$

and

$$\mu_{X,P} = \mathbb{E}_P[\varphi_k(X)], \qquad \mu_{\xi,P} = \mathbb{E}_P[\varphi_l\{\xi_P(W,Y)\}].$$

For $j = 1, \ldots, d_y$, define

$$v_{P,j}(w) \doteq \mathbb{E}_P[(\partial_j \varphi_l)\{\xi_P(W,Y)\} \mid W = w] \in \mathcal{H}_l, \qquad v_P = (v_{P,1}, \ldots, v_{P,d_y}).$$

Set

$$\eta_P = (m_P, v_P, \mu_{X,P}, \mu_{\xi,P}).$$

For $\eta = (m, v, \mu_X, \mu_\xi)$, define

$$D_\eta(x,w,y) \doteq \{\varphi_k(x) - \mu_X\} \otimes \left[\varphi_l\{y - m(w)\} - \mu_\xi - \sum_{j=1}^{d_y}\{y_j - m_j(w)\}v_j(w)\right].$$

Under Assumption 1, the EIF in Theorem 1 satisfies

$$\phi_0(o) = D_{\eta_0}(o) - \Psi_0.$$

Let I_1, I_2 be an equal split of $[n]$, and define

$$\mathbb{P}_{n,r} \doteq \frac{2}{n}\sum_{i \in I_r} \delta_{o_i}, \qquad r = 1, 2.$$

Let

$$\widehat{\eta}_r = (\widehat{m}_r, \widehat{v}_r, \widehat{\mu}_{X,r}, \widehat{\mu}_{\xi,r}),$$

be fitted on I_r, where

$$\widehat{\mu}_{X,r} = \mathbb{P}_{n,r}\varphi_k(X), \qquad \widehat{\mu}_{\xi,r} = \mathbb{P}_{n,r}\varphi_l\{Y - \widehat{m}_r(W)\}.$$

The two-fold cross-fitted estimator is

$$\check{\Psi}_n \doteq \frac{1}{2}\{\mathbb{P}_{n,1}D_{\widehat{\eta}_2} + \mathbb{P}_{n,2}D_{\widehat{\eta}_1}\}.$$

Lemma 3 (Foldwise L^2-consistency of the estimated non-centered EIF)**.** *Suppose Assumptions 1 to 4 and requirement 1A hold. Let $\widehat{\eta}_r = (\widehat{m}_r, \widehat{v}_r, \widehat{\mu}_{X,r}, \widehat{\mu}_{\xi,r})$, $r = 1, 2$, and define*

$$D_\eta(x,w,y) \doteq \{\varphi_k(x) - \mu_X\} \otimes \left[\varphi_l\{y - m(w)\} - \mu_\xi - \sum_{j=1}^{d_y}\{y_j - m_j(w)\}v_j(w)\right].$$

If, for $r = 1, 2$,

$$\|\widehat{m}_r - m_0\|_{L^\infty(P_{0,W})} + \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})} + \|\widehat{v}_r - v_0\|_{L^2(P_{0,W})} = o_p(1),$$

and the residual mean embedding satisfies

$$\|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} = o_p(1),$$

then

$$\|\widehat{\mu}_{X,r} - \mu_{X,0}\|_{\mathcal{H}_k} = O_p(n^{-1/2})$$

and

$$\|D_{\widehat{\eta}_r} - D_{\eta_0}\|_{L^2(P_0;\mathcal{H}_{kl})} = o_p(1).$$

Proof. The bound for $\widehat{\mu}_{X,r}$ follows from Assumption 2. We write

$$A_\eta(X) \doteq \varphi_k(X) - \mu_X, \qquad M_\eta(O) \doteq \varphi_l\{Y - m(W)\} - \mu_\xi - \sum_{j=1}^{d_y}\{Y_j - m_j(W)\}v_j(W).$$

Then

$$D_{\widehat{\eta}_r} - D_{\eta_0} = (A_{\widehat{\eta}_r} - A_{\eta_0}) \otimes M_{\eta_0} + A_{\widehat{\eta}_r} \otimes (M_{\widehat{\eta}_r} - M_{\eta_0}).$$

By Assumptions 2 and 3, requirement 1A, and lemma 1, $M_{\eta_0} \in L^2(P_0; \mathcal{H}_l)$. Therefore

$$\|(A_{\widehat{\eta}_r} - A_{\eta_0}) \otimes M_{\eta_0}\|_{L^2(P_0)} \lesssim \|\widehat{\mu}_{X,r} - \mu_{X,0}\|_{\mathcal{H}_k} = o_p(1).$$

Moreover, $\|A_{\widehat{\eta}_r}\|_{\mathcal{H}_k} \leq 2\kappa_k$, and

$$\begin{aligned}\|M_{\widehat{\eta}_r} - M_{\eta_0}\|_{L^2(P_0)} \lesssim \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})} + \|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} \\ + \|\xi_0^\top(\widehat{v}_r - v_0)\|_{L^2(P_0)} + \|(\widehat{m}_r - m_0)^\top \widehat{v}_r\|_{L^2(P_0)}.\end{aligned}$$

The third term is $o_p(1)$ by Requirement 1A and $\|\widehat{v}_r - v_0\|_{L^2(P_{0,W})} = o_p(1)$. For the last term, Lemma 1 and conditional Jensen's inequality give $v_0 \in L^\infty(P_{0,W}; \mathcal{H}_l^{d_y})$. Consistency of $\widehat{v}_r$ therefore gives $\|\widehat{v}_r\|_{L^2(P_{0,W})} = O_p(1)$, and

$$\|(\widehat{m}_r - m_0)^\top \widehat{v}_r\|_{L^2(P_0)} \leq \|\widehat{m}_r - m_0\|_{L^\infty(P_{0,W})} \|\widehat{v}_r\|_{L^2(P_{0,W})} = o_p(1).$$

Thus $\|M_{\widehat{\eta}_r} - M_{\eta_0}\|_{L^2(P_0)} = o_p(1)$, proving the claim. □

Lemma 4 (Upper bound for the remainder term in each fold)**.** *Suppose Assumptions 1 to 4 hold. For $\eta = (m, v, \mu_X, \mu_\xi)$, define*

$$R_0(\eta) \doteq P_0 D_\eta - \Psi_0.$$

Then, for $r = 1, 2$,

$$\begin{aligned}\|R_0(\widehat{\eta}_r)\|_{\mathcal{H}_{kl}} \lesssim \|\widehat{\mu}_{X,r} - \mu_{X,0}\|_{\mathcal{H}_k} \|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} \\ + \|\widehat{m}_r - m_0\|^2_{L^2(P_{0,W})} \\ + \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})} \|\widehat{v}_r - v_0\|_{L^2(P_{0,W})}.\end{aligned}$$

Consequently, under the conditions of Lemma 3, if

$$\|\widehat{m}_r - m_0\|^2_{L^2(P_{0,W})} + \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})} \|\widehat{v}_r - v_0\|_{L^2(P_{0,W})} = o_p(n^{-1/2}),$$

then $R_0(\widehat{\eta}_r) = o_p(n^{-1/2})$.

Proof. Let $\delta = m - m_0$, $\xi_0 = Y - m_0(W)$, and $A_{\mu_X}(W) = \varphi_k\{\pi(W)\} - \mu_X$. By Assumption 1, $A_{\mu_X}(W) = \varphi_k(X) - \mu_X$ almost surely. Using $\mathbb{E}_0[\xi_0 \mid W] = 0$,

$$\begin{aligned}R_0(\eta) = (\mu_{X,0} - \mu_X) \otimes (\mu_{\xi,0} - \mu_\xi) \\ + P_0\left[A_{\mu_X}(W) \otimes \{\varphi_l(\xi_0 - \delta(W)) - \varphi_l(\xi_0)\}\right] \\ + P_0\left[A_{\mu_X}(W) \otimes \sum_{j=1}^{d_y} \delta_j(W) v_j(W)\right].\end{aligned}$$

Add and subtract $\sum_{j=1}^{d_y} \delta_j(W)(\partial_j \varphi_l)(\xi_0)$. Since

$$v_{0,j}(W) = \mathbb{E}_0[(\partial_j \varphi_l)(\xi_0) \mid W],$$

the linear terms combine to

$$P_0\left[A_{\mu_X}(W) \otimes \sum_{j=1}^{d_y} \delta_j(W)\{v_j(W) - v_{0,j}(W)\}\right].$$

By the Banach-valued fundamental theorem of calculus and Assumption 4,

$$\begin{aligned}&\left\|\varphi_l(\xi_0 - \delta) - \varphi_l(\xi_0) + \sum_{j=1}^{d_y} \delta_j(\partial_j \varphi_l)(\xi_0)\right\|_{\mathcal{H}_l} \\ &= \left\|\int_0^1 \{D\varphi_l(\xi_0 - t\delta) - D\varphi_l(\xi_0)\}[\delta]\, dt\right\|_{\mathcal{H}_l} \\ &\leq \int_0^1 L_2 t \|\delta\|^2_{\mathbb{R}^{d_y}}\, dt = \frac{L_2}{2}\|\delta\|^2_{\mathbb{R}^{d_y}}.\end{aligned}$$

Since $\|A_{\mu_X}(W)\|_{\mathcal{H}_k} \le 2\kappa_k$, Cauchy-Schwarz gives

$$\begin{aligned}\|R_0(\eta)\|_{\mathcal{H}_{kl}} \lesssim\ & \|\mu_{X,0} - \mu_X\|_{\mathcal{H}_k}\|\mu_{\xi,0} - \mu_\xi\|_{\mathcal{H}_l} \\ & + \|m - m_0\|^2_{L^2(P_{0,W})} \\ & + \|m - m_0\|_{L^2(P_{0,W})}\|v - v_0\|_{L^2(P_{0,W})}.\end{aligned}$$

Taking $\eta = \widehat{\eta}_r$ gives the result. The final statement follows from $\|\widehat{\mu}_{X,r} - \mu_{X,0}\|_{\mathcal{H}_k} = O_p(n^{-1/2})$ and $\|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} = o_p(1)$, as shown in Lemma 3. □

Analysis of the cross-fitted one-step estimator.

Theorem 3 (Two-fold cross-fitted asymptotic linearity)**.** *Let $o_1, \dots, o_n \sim P_0^n$, and let $\check{\Psi}_n$ be defined above. Suppose Assumptions 1 to 4 and requirement 1A hold. In place of Requirements 1B and 1C, assume that, for $r = 1, 2$,*

$$\|\widehat{m}_r - m_0\|_{L^\infty(P_{0,W})} + \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})} + \|\widehat{v}_r - v_0\|_{L^2(P_{0,W})} = o_p(1),$$

and

$$\|\widehat{m}_r - m_0\|^2_{L^2(P_{0,W})} + \|\widehat{m}_r - m_0\|_{L^2(P_{0,W})}\|\widehat{v}_r - v_0\|_{L^2(P_{0,W})} = o_p(n^{-1/2}).$$

Also assume

$$\|\widehat{\mu}_{\xi,r} - \mu_{\xi,0}\|_{\mathcal{H}_l} = o_p(1), \qquad r = 1, 2.$$

Then

$$\check{\Psi}_n - \Psi_0 = \mathbb{P}_n\phi_0 + o_p(n^{-1/2}),$$

where $\phi_0 = D_{\eta_0} - \Psi_0$ is the EIF in Theorem 1. Hence

$$\sqrt{n}(\check{\Psi}_n - \Psi_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_0(o_i) + o_p(1) \rightsquigarrow \mathbb{H} \quad \text{in } \mathcal{H}_{kl},$$

where $\mathbb{H}$ is a mean-zero Gaussian element in $\mathcal{H}_{kl}$ with

$$\mathrm{Var}\{\langle \mathbb{H}, h\rangle_{\mathcal{H}_{kl}}\} = \mathbb{E}_0[\langle \phi_0(O), h\rangle^2_{\mathcal{H}_{kl}}], \qquad h \in \mathcal{H}_{kl}.$$

Proof of Theorem 3. Set $R_0(\eta) \doteq P_0 D_\eta - \Psi_0$. Since $D_{\eta_0} - \Psi_0 = \phi_0$,

$$\begin{aligned}\check{\Psi}_n - \Psi_0 = \ & \mathbb{P}_n\phi_0 \\ & + \frac{1}{2}(\mathbb{P}_{n,1} - P_0)(D_{\widehat{\eta}_2} - D_{\eta_0}) + \frac{1}{2}(\mathbb{P}_{n,2} - P_0)(D_{\widehat{\eta}_1} - D_{\eta_0}) \\ & + \frac{1}{2}\{R_0(\widehat{\eta}_1) + R_0(\widehat{\eta}_2)\}.\end{aligned}$$

By Lemma 3 and sample splitting, we can control the stochastic equicontinuity term as follows

$$\begin{aligned}&\frac{1}{2}(\mathbb{P}_{n,1} - P_0)(D_{\widehat{\eta}_2} - D_{\eta_0}) + \frac{1}{2}(\mathbb{P}_{n,2} - P_0)(D_{\widehat{\eta}_1} - D_{\eta_0}) \\ &\quad = o_p(n^{-1/2}).\end{aligned}$$

By Lemma 4,

$$R_0(\widehat{\eta}_1) + R_0(\widehat{\eta}_2) = o_p(n^{-1/2}).$$

Therefore

$$\check{\Psi}_n - \Psi_0 = \mathbb{P}_n\phi_0 + o_p(n^{-1/2}).$$

The weak convergence to $\mathbb{H}$ follows from applying a Hilbert-valued CLT, using the fact that $\mathcal{H}_{kl}$ is separable and $\phi_0 \in L^2_0(P_0; \mathcal{H}_{kl})$. □

E Convergence rates for the nuisance estimators

We exhibit elementary sufficient conditions for the conditions of Theorem 3 using general purpose Hölder smoothness assumptions. For concreteness, we focus on kernel ridge regression (KRR) and vector-valued kernel ridge regression (vvKRR) estimators, due to their ease of implementation, considerable flexibility with the choice of input domain, and strong theoretical guarantees for both scalar and infinite-dimensional Hilbert-valued outputs. However, we emphasize that leveraging the von Mises expansion for the bias of a plug-in estimator, our method can accommodate any flexible machine learning estimators, such as random forests or neural networks, that a practitioner may choose to deploy due to the specific nature of a dataset, and yield a debiased estimator where the plug-in bias enters only in second order. We further remark that it may be of strong theoretical interest to analyze the effect of a structure-agnostic expansion of our statistical estimand, and analyze both inference optimal choice of hyperparameter selection in our setting, as well as the effect of performing further sample splitting to estimate our primary nuisance m and secondary (vector-valued) nuisance v on different, hence independent, splits of the sample. Prior work has studied the expected conditional covariance (ECC) estimand [McClean et al., 2026, McGrath and Mukherjee, 2026], which arises in conditional independence testing. Such an analysis is beyond the scope of our work, but offers an intriguing possibility towards further weakening the minimal smoothness requirements for the remainder in the Von Mises expansion to be asymptotically negligible. Our analysis makes heavy use of existing machinery developed in Fischer and Steinwart [2020] for Sobolev norm KRR and Li et al. [2024] for Sobolev norm vvKRR, as well as a long line of prior work such as Blanchard and Mücke [2018], Lin and Cevher [2018], Lin et al. [2020].

We consider a general i.i.d. sample I used to fit ***both*** v and m. Let $N = |I|$ be its size and let $P^I_{W,N} = N^{-1}\sum_{i\in I}\delta_{W_i}$ denote its empirical covariate distribution. This sample can then be instantiated as the fold-wise fitting sample in the two-fold sample splitting scheme in the main text or the general K-fold cross fitting.

v-estimator. We first study the vector-valued regression problem. We introduce an additional reproducing kernel $q : \mathcal{W}\times\mathcal{W}\to\mathbb{R}$, which we assume to be measurable and to satisfy $\sup_{w\in\mathcal{W}} q(w,w)\le\kappa_q^2$ for a finite constant $\kappa_q\ge 0$. We emphasize that the purpose of this kernel is to define a notion of similarity on $\mathcal{W}$ for the sole purpose of defining a nonparametric regression from $\mathcal{W}$ to a general output space, unlike l whose feature map φ_l is used to featurize residuals in $\mathbb{R}^{d_y}$ for the purpose of defining the kernelized test statistic. Let $\mathcal{G}$ denote the vector-valued RKHS induced by the ***separable*** vector-valued kernel $Q \doteq q\mathrm{Id}_{\mathcal{H}_l} : \mathcal{W}\times\mathcal{W}\to\mathcal{B}(\mathcal{H}_l)$, which satisfies, for $w, w'\in\mathcal{W}$,

$$Q(w,w') = q(w,w')\cdot \mathrm{Id}_{\mathcal{H}_l}.$$

Therefore our vector-valued kernel does not encode interaction between different directions of $\mathcal{H}_l$. We further write $\varphi_q(w)\doteq q(\cdot,w)$ for $w\in\mathcal{W}$ as the canonical feature map of q. For $j\in[d_y]$ and a regularization parameter $\lambda>0$, we define the following optimisation problems,

$$\widehat{v}_{j,\lambda} = \operatorname*{argmin}_{v\in\mathcal{G}}\left\{\frac{1}{N}\sum_{i\in I}\|\partial_j\varphi_l(Y_i-\widehat{m}(W_i))-v(W_i)\|^2_{\mathcal{H}_l}+\lambda\|v\|^2_{\mathcal{G}}\right\}, \tag{8}$$

$$\widetilde{v}_{j,\lambda} = \operatorname*{argmin}_{v\in\mathcal{G}}\left\{\frac{1}{N}\sum_{i\in I}\|\partial_j\varphi_l(Y_i-m_0(W_i))-v(W_i)\|^2_{\mathcal{H}_l}+\lambda\|v\|^2_{\mathcal{G}}\right\}. \tag{9}$$

Notice that the second estimator Eq. (9) is ***infeasible***, as we do not have access to m_0. Our proof strategy is as follows. We will use a generic perturbation analysis to bound the error of $\widehat{v}_{j,\lambda}-\widetilde{v}_{j,\lambda}$ in terms of the MSE of $\widehat{m}-m_0$, and then bound the error of $\widetilde{v}_{j,\lambda}-v_{0,j}$. The latter step can be tackled using either an empirical process argument or an integral operator based approach. In our analysis we adopt the latter approach and employ an operator-valued Bernstein inequality to establish our guarantees [Minsker, 2017, Fischer and Steinwart, 2020].

Step I. v-estimator perturbation bound. We define the population and empirical covariance operators associated to q as follows:

$$C_q \doteq \mathbb{E}_0[\varphi_q(W) \otimes \varphi_q(W)],$$
$$\widehat{C}_q \doteq \frac{1}{N}\sum_{i\in I} \varphi_q(W_i) \otimes \varphi_q(W_i).$$

Both operators act on the scalar reproducing kernel Hilbert space $\mathcal{H}_q$ of q. Here $(a\otimes b)h = a\langle b, h\rangle_{\mathcal{H}_q}$ for $a, b, h \in \mathcal{H}_q$. For two self-adjoint operators A, B, we use the Löwner order notation $A \preceq B$ to denote that $B - A$ is a positive semidefinite operator.

Lemma 5 (Error of empirical v-estimator relative to the oracle v-estimator)**.** *Suppose q is bounded and measurable and l satisfies Assumption 4. For $j \in \{1, \ldots, d_y\}$ and $\lambda > 0$, let $\widehat{v}_{j,\lambda}$ and $\widetilde{v}_{j,\lambda}$ be the estimators in (8) and (9). On the event*

$$\left\|(C_q + \lambda \,\mathrm{Id})^{-1/2}(C_q - \widehat{C}_q)(C_q + \lambda \,\mathrm{Id})^{-1/2}\right\|_{\mathrm{op}} \le \frac{1}{2}, \tag{10}$$

we have

$$\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|_{L^2(P_{0,W};\mathcal{H}_l)} \le \sqrt{2}L_2\|\widehat{m} - m_0\|_{L^2(P^I_{W,N};\mathbb{R}^{d_y})}. \tag{11}$$

Proof. We first derive the normal equations. For labels $\zeta = (\zeta_i)_{i\in I} \in \mathcal{H}_l^N$, define the ridge objective $\mathcal{L}_\lambda(\cdot;\zeta) : \mathcal{G} \to \mathbb{R}$ by

$$\mathcal{L}_\lambda(f;\zeta) = \frac{1}{N}\sum_{i\in I}\|f(W_i) - \zeta_i\|^2_{\mathcal{H}_l} + \lambda\|f\|^2_{\mathcal{G}}.$$

The reproducing property and the Cauchy–Schwarz inequality give

$$\begin{aligned}\|f(w)\|_{\mathcal{H}_l} &= \sup_{\|u\|_{\mathcal{H}_l}\le 1} |\langle f, Q(\cdot, w)u\rangle_{\mathcal{G}}| \\ &\le \|f\|_{\mathcal{G}} \sup_{\|u\|_{\mathcal{H}_l}\le 1} \|Q(\cdot, w)u\|_{\mathcal{G}} \\ &\le \sqrt{q(w,w)}\,\|f\|_{\mathcal{G}} \le \kappa_q\|f\|_{\mathcal{G}}.\end{aligned}$$

Thus evaluation is a bounded linear map. The objective is weakly lower semicontinuous, and $\mathcal{L}_\lambda(f;\zeta) \ge \lambda\|f\|^2_{\mathcal{G}}$. The direct method in a Hilbert space gives a minimizer. The positive quadratic term $\lambda\|f\|^2_{\mathcal{G}}$ makes the objective strictly convex, so the minimizer is unique.

For $f, g \in \mathcal{G}$ and $t \in \mathbb{R}$, expansion of both squared norms gives

$$\begin{aligned}\mathcal{L}_\lambda(f + tg;\zeta) = \frac{1}{N}\sum_{i\in I}\Big\{&\|f(W_i) - \zeta_i\|^2_{\mathcal{H}_l} \\ &+ 2t\langle f(W_i) - \zeta_i, g(W_i)\rangle_{\mathcal{H}_l} + t^2\|g(W_i)\|^2_{\mathcal{H}_l}\Big\} \\ &+ \lambda\|f\|^2_{\mathcal{G}} + 2\lambda t\langle f, g\rangle_{\mathcal{G}} + \lambda t^2\|g\|^2_{\mathcal{G}} \\ = \mathcal{L}_\lambda(f;\zeta) &+ 2t\left\{\frac{1}{N}\sum_{i\in I}\langle f(W_i) - \zeta_i, g(W_i)\rangle_{\mathcal{H}_l} + \lambda\langle f, g\rangle_{\mathcal{G}}\right\} \\ &+ t^2\left\{\frac{1}{N}\sum_{i\in I}\|g(W_i)\|^2_{\mathcal{H}_l} + \lambda\|g\|^2_{\mathcal{G}}\right\}.\end{aligned}$$

Differentiating this polynomial in t at zero yields

$$\frac{d}{dt}\mathcal{L}_\lambda(f + tg;\zeta)\Big|_{t=0} = 2\left\{\frac{1}{N}\sum_{i\in I}\langle f(W_i) - \zeta_i, g(W_i)\rangle_{\mathcal{H}_l} + \lambda\langle f, g\rangle_{\mathcal{G}}\right\}.$$

At the minimizer, this derivative is zero for every $g \in \mathcal{G}$. Substituting each set of labels therefore gives the two equations

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in I}\langle \widehat{v}_{j,\lambda}(W_i) - \partial_j\varphi_l\{Y_i - \widehat{m}(W_i)\}, g(W_i)\rangle_{\mathcal{H}_l} \\
&\qquad + \lambda\langle \widehat{v}_{j,\lambda}, g\rangle_{\mathcal{G}} = 0, \\
&\frac{1}{N}\sum_{i\in I}\langle \widetilde{v}_{j,\lambda}(W_i) - \partial_j\varphi_l\{Y_i - m_0(W_i)\}, g(W_i)\rangle_{\mathcal{H}_l} \\
&\qquad + \lambda\langle \widetilde{v}_{j,\lambda}, g\rangle_{\mathcal{G}} = 0.
\end{aligned}$$

Subtracting the second equation from the first and moving the label difference to the right gives

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in I}\langle (\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda})(W_i), g(W_i)\rangle_{\mathcal{H}_l} \\
&\qquad + \lambda\langle \widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}, g\rangle_{\mathcal{G}} \\
&= \frac{1}{N}\sum_{i\in I}\langle \partial_j\varphi_l\{Y_i - \widehat{m}(W_i)\} - \partial_j\varphi_l\{Y_i - m_0(W_i)\}, g(W_i)\rangle_{\mathcal{H}_l}.
\end{aligned}$$

We choose $g = \widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}$ to obtain

$$\begin{aligned}
&\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{L^2(P^I_{W,N};\mathcal{H}_l)} + \lambda\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{\mathcal{G}} \\
&= \frac{1}{N}\sum_{i\in I}\langle \partial_j\varphi_l\{Y_i - \widehat{m}(W_i)\} - \partial_j\varphi_l\{Y_i - m_0(W_i)\}, (\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda})(W_i)\rangle_{\mathcal{H}_l}.
\end{aligned}$$

Let $e_j \in \mathbb{R}^{d_y}$ be the jth standard coordinate vector. We have that $\partial_j\varphi_l(z) = D\varphi_l(z)[e_j]$. We apply Assumption 4 to obtain

$$\begin{aligned}
&\|\partial_j\varphi_l\{Y_i - \widehat{m}(W_i)\} - \partial_j\varphi_l\{Y_i - m_0(W_i)\}\|_{\mathcal{H}_l} \\
&= \big\|\big[D\varphi_l\{Y_i - \widehat{m}(W_i)\} - D\varphi_l\{Y_i - m_0(W_i)\}\big][e_j]\big\|_{\mathcal{H}_l} \\
&\le \|D\varphi_l\{Y_i - \widehat{m}(W_i)\} - D\varphi_l\{Y_i - m_0(W_i)\}\|_{\mathrm{op}}\|e_j\|_{\mathbb{R}^{d_y}} \\
&\le L_2\|\widehat{m}(W_i) - m_0(W_i)\|_{\mathbb{R}^{d_y}}.
\end{aligned}$$

Applying the Cauchy–Schwarz inequality yields

$$\begin{aligned}
&\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{L^2(P^I_{W,N};\mathcal{H}_l)} + \lambda\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{\mathcal{G}} \\
&\le \frac{L_2}{N}\sum_{i\in I}\|\widehat{m}(W_i) - m_0(W_i)\|_{\mathbb{R}^{d_y}}\|(\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda})(W_i)\|_{\mathcal{H}_l} \\
&\le L_2\left\{\frac{1}{N}\sum_{i\in I}\|\widehat{m}(W_i) - m_0(W_i)\|^2_{\mathbb{R}^{d_y}}\right\}^{1/2}\left\{\frac{1}{N}\sum_{i\in I}\|(\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda})(W_i)\|^2_{\mathcal{H}_l}\right\}^{1/2} \\
&= L_2\|\widehat{m} - m_0\|_{L^2(P^I_{W,N};\mathbb{R}^{d_y})}\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|_{L^2(P^I_{W,N};\mathcal{H}_l)} \\
&\le \frac{L_2^2}{2}\|\widehat{m} - m_0\|^2_{L^2(P^I_{W,N};\mathbb{R}^{d_y})} + \frac{1}{2}\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{L^2(P^I_{W,N};\mathcal{H}_l)},
\end{aligned}$$

where the last step follows from the inequality $2ab \le a^2 + b^2$. We rearrange to obtain

$$\begin{aligned}
&\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{L^2(P^I_{W,N};\mathcal{H}_l)} + 2\lambda\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{\mathcal{G}} \\
&\qquad \le L_2^2\|\widehat{m} - m_0\|^2_{L^2(P^I_{W,N};\mathbb{R}^{d_y})}.
\end{aligned}$$

We now use (10) to bound the population L^2 norm. Measurability of q and separability of $\mathcal{H}_q$ imply that φ_q is strongly measurable. The identity

$$\|\varphi_q(w)\otimes\varphi_q(w)\|_{\mathrm{op}} = \|\varphi_q(w)\|^2_{\mathcal{H}_q} = q(w,w) \le \kappa_q^2$$

therefore makes the covariance expectation well defined. For $h \in \mathcal{H}_q$, the reproducing property gives

$$\langle h, C_q h\rangle_{\mathcal{H}_q} = \mathbb{E}_0[\langle h, \varphi_q(W)\rangle^2_{\mathcal{H}_q}] = \|h\|^2_{L^2(P_{0,W};\mathbb{R})},$$
$$\langle h, \widehat{C}_q h\rangle_{\mathcal{H}_q} = \frac{1}{N}\sum_{i\in I}\langle h, \varphi_q(W_i)\rangle^2_{\mathcal{H}_q} = \|h\|^2_{L^2(P^I_{W,N};\mathbb{R})}.$$

The scalar reproducing bound further gives

$$0 \le \langle h, C_q h\rangle_{\mathcal{H}_q} \le \mathbb{E}_0[q(W,W)]\|h\|^2_{\mathcal{H}_q} \le \kappa_q^2\|h\|^2_{\mathcal{H}_q},$$
$$0 \le \langle h, \widehat{C}_q h\rangle_{\mathcal{H}_q} \le \left\{\frac{1}{N}\sum_{i\in I} q(W_i, W_i)\right\}\|h\|^2_{\mathcal{H}_q} \le \kappa_q^2\|h\|^2_{\mathcal{H}_q}.$$

Thus both covariance operators are positive and bounded. In particular, $C_q + \lambda \operatorname{Id}_{\mathcal{H}_q}$ has a bounded inverse and a positive square root by the spectral theorem.

On the event (10), for every $h \in \mathcal{H}_q$,

$$\begin{aligned}
&\left|\langle h, (C_q - \widehat{C}_q)h\rangle_{\mathcal{H}_q}\right| \\
&= \Big|\Big\langle (C_q + \lambda \operatorname{Id}_{\mathcal{H}_q})^{1/2} h, \\
&\qquad (C_q + \lambda \operatorname{Id}_{\mathcal{H}_q})^{-1/2}(C_q - \widehat{C}_q)(C_q + \lambda \operatorname{Id}_{\mathcal{H}_q})^{-1/2}(C_q + \lambda \operatorname{Id}_{\mathcal{H}_q})^{1/2} h\Big\rangle_{\mathcal{H}_q}\Big| \\
&\le \frac{1}{2}\|(C_q + \lambda \operatorname{Id}_{\mathcal{H}_q})^{1/2} h\|^2_{\mathcal{H}_q} \\
&= \frac{1}{2}\langle h, C_q h\rangle_{\mathcal{H}_q} + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}_q}.
\end{aligned}$$

Therefore

$$\langle h, C_q h\rangle_{\mathcal{H}_q} - \langle h, \widehat{C}_q h\rangle_{\mathcal{H}_q} \le \frac{1}{2}\langle h, C_q h\rangle_{\mathcal{H}_q} + \frac{\lambda}{2}\|h\|^2_{\mathcal{H}_q},$$
$$\langle h, C_q h\rangle_{\mathcal{H}_q} \le 2\langle h, \widehat{C}_q h\rangle_{\mathcal{H}_q} + \lambda\|h\|^2_{\mathcal{H}_q}.$$

Equivalently, in the Löwner order,

$$C_q \preceq 2\widehat{C}_q + \lambda \operatorname{Id}_{\mathcal{H}_q}.$$

We now transfer this inequality to $\mathcal{H}_l$-valued functions. Let $(h_b)_{b\in\mathcal{J}}$ be an orthonormal basis of $\mathcal{H}_l$, where $\mathcal{J}$ is finite or countably infinite. For $f \in \mathcal{G}$, its scalar coordinate function is $f_b(w) = \langle f(w), h_b\rangle_{\mathcal{H}_l}$. We first prove that these coordinate functions belong to $\mathcal{H}_q$ and satisfy

$$\sum_{b\in\mathcal{J}} \|f_b\|^2_{\mathcal{H}_q} = \|f\|^2_{\mathcal{G}}.$$

For a finite sum of kernel sections, with $S \ge 1$, $w_1, \dots, w_S \in \mathcal{W}$, and $u_1, \dots, u_S \in \mathcal{H}_l$,

$$f = \sum_{s=1}^{S} Q(\cdot, w_s)u_s = \sum_{s=1}^{S} q(\cdot, w_s)u_s,$$
$$f_b = \sum_{s=1}^{S}\langle u_s, h_b\rangle_{\mathcal{H}_l}\, q(\cdot, w_s) \in \mathcal{H}_q.$$

The reproducing properties in $\mathcal{H}_q$ and $\mathcal{G}$, followed by Parseval's identity in $\mathcal{H}_l$, give

$$\begin{aligned}\sum_{b\in\mathcal{J}}\|f_b\|_{\mathcal{H}_q}^2 &= \sum_{b\in\mathcal{J}}\sum_{s,t=1}^{S} q(w_s,w_t)\langle u_s,h_b\rangle_{\mathcal{H}_l}\langle u_t,h_b\rangle_{\mathcal{H}_l}\\ &= \sum_{s,t=1}^{S} q(w_s,w_t)\sum_{b\in\mathcal{J}}\langle u_s,h_b\rangle_{\mathcal{H}_l}\langle u_t,h_b\rangle_{\mathcal{H}_l}\\ &= \sum_{s,t=1}^{S} q(w_s,w_t)\langle u_s,u_t\rangle_{\mathcal{H}_l}\\ &= \left\|\sum_{s=1}^{S} Q(\cdot,w_s)u_s\right\|_{\mathcal{G}}^2 = \|f\|_{\mathcal{G}}^2.\end{aligned}$$

The interchange of sums is justified by the finite sums over s,t and the Cauchy–Schwarz bound

$$\begin{aligned}&\sum_{b\in\mathcal{J}}|\langle u_s,h_b\rangle_{\mathcal{H}_l}\langle u_t,h_b\rangle_{\mathcal{H}_l}|\\ &\quad\le\left(\sum_{b\in\mathcal{J}}|\langle u_s,h_b\rangle_{\mathcal{H}_l}|^2\right)^{1/2}\left(\sum_{b\in\mathcal{J}}|\langle u_t,h_b\rangle_{\mathcal{H}_l}|^2\right)^{1/2} = \|u_s\|_{\mathcal{H}_l}\|u_t\|_{\mathcal{H}_l}.\end{aligned}$$

For an arbitrary $f\in\mathcal{G}$, choose finite sums of kernel sections $f^{(a)}\in\mathcal{G}$, indexed by integers $a\ge 1$, such that $\|f^{(a)}-f\|_{\mathcal{G}}\to 0$. For each a and b, write $f_b^{(a)}(w)=\langle f^{(a)}(w),h_b\rangle_{\mathcal{H}_l}$. The identity for finite sums implies

$$\sum_{b\in\mathcal{J}}\|f_b^{(a)}-f_b^{(c)}\|_{\mathcal{H}_q}^2 = \|f^{(a)}-f^{(c)}\|_{\mathcal{G}}^2 \longrightarrow 0 \qquad (a,c\longrightarrow\infty).$$

Thus the coordinate sequences converge in the complete Hilbert space $\ell^2(\mathcal{J};\mathcal{H}_q)$ of square-summable $\mathcal{H}_q$-valued sequences. Continuity of scalar evaluation and the bound

$$\begin{aligned}|f_b^{(a)}(w)-\langle f(w),h_b\rangle_{\mathcal{H}_l}| &\le \|f^{(a)}(w)-f(w)\|_{\mathcal{H}_l}\\ &\le \kappa_q\|f^{(a)}-f\|_{\mathcal{G}}\longrightarrow 0\end{aligned}$$

identify its bth limit as the function f_b. Continuity of the norms in these Hilbert spaces now gives

$$\sum_{b\in\mathcal{J}}\|f_b\|_{\mathcal{H}_q}^2 = \lim_{a\to\infty}\sum_{b\in\mathcal{J}}\|f_b^{(a)}\|_{\mathcal{H}_q}^2 = \lim_{a\to\infty}\|f^{(a)}\|_{\mathcal{G}}^2 = \|f\|_{\mathcal{G}}^2.$$

Pointwise Parseval's identity, Tonelli's theorem for the nonnegative coordinate squares, and the scalar covariance identities give

$$\begin{aligned}\|f\|_{L^2(P_{0,W};\mathcal{H}_l)}^2 &= \mathbb{E}_0\left[\sum_{b\in\mathcal{J}}|f_b(W)|^2\right]\\ &= \sum_{b\in\mathcal{J}}\mathbb{E}_0[|f_b(W)|^2] = \sum_{b\in\mathcal{J}}\langle f_b,C_q f_b\rangle_{\mathcal{H}_q},\\ \|f\|_{L^2(P_{W,N}^I;\mathcal{H}_l)}^2 &= \frac{1}{N}\sum_{i\in I}\sum_{b\in\mathcal{J}}|f_b(W_i)|^2\\ &= \sum_{b\in\mathcal{J}}\frac{1}{N}\sum_{i\in I}|f_b(W_i)|^2 = \sum_{b\in\mathcal{J}}\langle f_b,\widehat{C}_q f_b\rangle_{\mathcal{H}_q}.\end{aligned}$$

Applying the scalar covariance comparison to each f_b and summing therefore yields

$$\begin{aligned}\|f\|^2_{L^2(P_{0,W};\mathcal{H}_l)} &= \sum_{b\in\mathcal{J}} \langle f_b, C_q f_b\rangle_{\mathcal{H}_q} \\ &\le \sum_{b\in\mathcal{J}} \left\{ 2\langle f_b, \widehat{C}_q f_b\rangle_{\mathcal{H}_q} + \lambda \|f_b\|^2_{\mathcal{H}_q} \right\} \\ &= 2\|f\|^2_{L^2(P^I_{W,N};\mathcal{H}_l)} + \lambda \|f\|^2_{\mathcal{G}}.\end{aligned}$$

We apply this inequality to $f = \widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}$ and combine it with the bound for the empirical L^2 norm above:

$$\begin{aligned}&\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{L^2(P_{0,W};\mathcal{H}_l)} \\ &\le 2\|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{L^2(P^I_{W,N};\mathcal{H}_l)} + \lambda \|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{\mathcal{G}} \\ &\le 2\left\{ \|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{L^2(P^I_{W,N};\mathcal{H}_l)} + 2\lambda \|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|^2_{\mathcal{G}} \right\} \\ &\le 2L_2^2 \|\widehat{m} - m_0\|^2_{L^2(P^I_{W,N};\mathbb{R}^{d_y})}.\end{aligned}$$

Taking square roots proves the population bound. □

Step II. Convergence of the v-estimator.

Corollary 1 (Convergence of the v-estimator)**.** *Suppose the assumptions of Lemma 5 hold. Let I contain N independent observations from P_0, and let $\lambda_N > 0$ be a deterministic sequence for which (10) holds with probability tending to one. For a coordinate $j \in \{1, \dots, d_y\}$, let*

$$v_{0,j}(w) = \mathbb{E}_0[\partial_j \varphi_l\{Y - m_0(W)\} \mid W = w].$$

If, for deterministic positive sequences $a_N, b_N \to 0$,

$$\begin{aligned}\|\widehat{m} - m_0\|_{L^2(P^I_{W,N};\mathbb{R}^{d_y})} &= O_p(a_N), \\ \|\widetilde{v}_{j,\lambda_N} - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} &= O_p(b_N),\end{aligned}$$

then

$$\|\widehat{v}_{j,\lambda_N} - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} = O_p(a_N + b_N). \tag{12}$$

Proof. By the triangle inequality and Lemma 5, on the event (10),

$$\begin{aligned}\|\widehat{v}_{j,\lambda} - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} &\le \|\widehat{v}_{j,\lambda} - \widetilde{v}_{j,\lambda}\|_{L^2(P_{0,W};\mathcal{H}_l)} \\ &\quad + \|\widetilde{v}_{j,\lambda} - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} \\ &\le \sqrt{2} L_2 \|\widehat{m} - m_0\|_{L^2(P^I_{W,N};\mathbb{R}^{d_y})} \\ &\quad + \|\widetilde{v}_{j,\lambda} - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)}.\end{aligned}$$

For any $M > 0$, the deterministic inequality and the union bound imply

$$\begin{aligned}&\mathbb{P}\Big(\|\widehat{v}_{j,\lambda_N} - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} > M(\sqrt{2}L_2 a_N + b_N) \Big) \\ &\le \mathbb{P}\left(\left\| (C_q + \lambda_N \operatorname{Id}_{\mathcal{H}_q})^{-1/2} (C_q - \widehat{C}_q)(C_q + \lambda_N \operatorname{Id}_{\mathcal{H}_q})^{-1/2} \right\|_{\mathrm{op}} > \frac{1}{2} \right) \\ &\quad + \mathbb{P}\Big(\|\widehat{m} - m_0\|_{L^2(P^I_{W,N};\mathbb{R}^{d_y})} > M a_N \Big) \\ &\quad + \mathbb{P}\big(\|\widetilde{v}_{j,\lambda_N} - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} > M b_N \big).\end{aligned}$$

For every $\varepsilon > 0$, the two assumed bounds in probability provide a finite M for which the last two probabilities are each at most $\varepsilon/3$ for all sufficiently large N. The first probability is also at most $\varepsilon/3$ for all sufficiently large N by the assumption on (10). Since $\sqrt{2}L_2 a_N + b_N \le \max\{\sqrt{2}L_2, 1\}(a_N + b_N)$, we obtain

$$\|\widehat{v}_{j,\lambda_N} - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} = O_p(a_N + b_N).$$

□

The bound in (12) uses the error of $\widehat{m}$ on the observations used to fit $\widehat{v}_{j,\lambda_N}$. The Sobolev example below bounds this empirical error when both estimators use the same observations.

E.1 The residual mean embedding and sufficient nuisance rates

Lemma 6 (Convergence of the residual mean embedding)**.** *Let I contain N independent observations from P_0. Suppose $\sup_u l(u,u) \le \kappa_l^2$ and l satisfies Assumption 4. For any measurable regression estimator $\widehat{m}$, define*

$$\widehat{\mu}_\xi^I = \frac{1}{N}\sum_{i\in I} \varphi_l\{Y_i - \widehat{m}(W_i)\}, \qquad \mu_{\xi,0} = \mathbb{E}_0[\varphi_l\{Y - m_0(W)\}].$$

Then

$$\begin{aligned}\|\widehat{\mu}_\xi^I - \mu_{\xi,0}\|_{\mathcal{H}_l} \le \sqrt{2\kappa_l L_2}\|\widehat{m} - m_0\|_{L^2(P^I_{W,N};\mathbb{R}^{d_y})} \\ + \left\|\frac{1}{N}\sum_{i\in I}[\varphi_l\{Y_i - m_0(W_i)\} - \mu_{\xi,0}]\right\|_{\mathcal{H}_l},\end{aligned} \tag{13}$$

and the second term is $O_p(N^{-1/2})$.

Proof. We add and subtract the empirical mean computed with m_0. The triangle inequality gives

$$\begin{aligned}\|\widehat{\mu}_\xi^I - \mu_{\xi,0}\|_{\mathcal{H}_l} \le \frac{1}{N}\sum_{i\in I}\|\varphi_l\{Y_i - \widehat{m}(W_i)\} - \varphi_l\{Y_i - m_0(W_i)\}\|_{\mathcal{H}_l} \\ + \left\|\frac{1}{N}\sum_{i\in I}[\varphi_l\{Y_i - m_0(W_i)\} - \mu_{\xi,0}]\right\|_{\mathcal{H}_l}.\end{aligned}$$

For the first term, Lemma 1 and the Cauchy–Schwarz inequality yield

$$\begin{aligned}&\frac{1}{N}\sum_{i\in I}\|\varphi_l\{Y_i - \widehat{m}(W_i)\} - \varphi_l\{Y_i - m_0(W_i)\}\|_{\mathcal{H}_l} \\ &\quad\le \frac{\sqrt{2\kappa_l L_2}}{N}\sum_{i\in I}\|\widehat{m}(W_i) - m_0(W_i)\|_{\mathbb{R}^{d_y}} \\ &\quad\le \sqrt{2\kappa_l L_2}\left\{\frac{1}{N}\sum_{i\in I}\|\widehat{m}(W_i) - m_0(W_i)\|^2_{\mathbb{R}^{d_y}}\right\}^{1/2}.\end{aligned}$$

For the second term, independence and centering remove the terms with distinct observation indices:

$$\begin{aligned}&\mathbb{E}_0\left\|\frac{1}{N}\sum_{i\in I}[\varphi_l\{Y_i - m_0(W_i)\} - \mu_{\xi,0}]\right\|^2_{\mathcal{H}_l} \\ &= \frac{1}{N^2}\sum_{i\in I}\mathbb{E}_0\|\varphi_l\{Y_i - m_0(W_i)\} - \mu_{\xi,0}\|^2_{\mathcal{H}_l} \\ &= \frac{1}{N}\left(\mathbb{E}_0\|\varphi_l\{Y - m_0(W)\}\|^2_{\mathcal{H}_l} - \|\mu_{\xi,0}\|^2_{\mathcal{H}_l}\right) \le \frac{\kappa_l^2}{N}.\end{aligned}$$

Consequently, for every $M > 0$, Markov's inequality gives

$$\mathbb{P}_0\left(\left\|\frac{1}{N}\sum_{i\in I}[\varphi_l\{Y_i - m_0(W_i)\} - \mu_{\xi,0}]\right\|_{\mathcal{H}_l} > \frac{M}{\sqrt{N}}\right) \le \frac{\kappa_l^2}{M^2}.$$

This proves the bound in probability. □

Corollary 2 (Sufficient rates for asymptotic linearity)**.** *Suppose the model, kernel, and conditional moment assumptions of Theorem 3 hold. Let K be fixed, with equal evaluation folds and training sizes proportional to n. For a training sample of size N, use a deterministic regularization parameter $\lambda_N > 0$. In every training fold, suppose (10) holds with probability tending to one and, for each $j \in \{1, \dots, d_y\}$,*

$$\|\widetilde{v}_{j,\lambda_N} - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} = O_p(n^{-\gamma_v}),$$

where $\gamma_v > 0$. Suppose that, in every training fold, the first-stage estimator satisfies

$$\|\widehat{m} - m_0\|_{L^2(P_{0,W};\mathbb{R}^{d_y})} + \|\widehat{m} - m_0\|_{L^2(P^I_{W,N};\mathbb{R}^{d_y})} = O_p(n^{-\gamma_m}),$$
$$\|\widehat{m} - m_0\|_{L^\infty(P_{0,W};\mathbb{R}^{d_y})} = o_p(1),$$

for a constant $\gamma_m > 0$. Compute the residual mean embedding on I and the covariate mean embedding on a training sample of independent observations of size proportional to n. If

$$\gamma_m > \frac{1}{4}, \qquad \gamma_m + \gamma_v > \frac{1}{2}, \tag{14}$$

then the cross-fitted one-step estimator satisfies

$$\sqrt{n}(\check{\Psi}_n - \Psi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_0(O_i) + o_p(1) \rightsquigarrow \mathbb{H} \quad \text{in } \mathcal{H}_{kl},$$

where ϕ_0 is the efficient influence function in Theorem 1 and $\mathbb{H}$ is its centered Gaussian limit.

Proof. By Corollary 1, in every training fold,

$$\|\widehat{v}_j - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} = O_p(n^{-\gamma_m} + n^{-\gamma_v}).$$

Since d_y is fixed, the same rate holds for the vector $v = (v_1, \dots, v_{d_y})$. Therefore

$$\begin{aligned}
&\|\widehat{m} - m_0\|^2_{L^2(P_{0,W})} + \|\widehat{m} - m_0\|_{L^2(P_{0,W})}\|\widehat{v} - v_0\|_{L^2(P_{0,W})} \\
&= O_p(n^{-2\gamma_m}) + O_p(n^{-\gamma_m})O_p(n^{-\gamma_m} + n^{-\gamma_v}) \\
&= O_p(n^{-2\gamma_m} + n^{-(\gamma_m+\gamma_v)}) = o_p(n^{-1/2}),
\end{aligned}$$

where the final equality uses (14). By Lemma 6 and boundedness of k,

$$\|\widehat{\mu}_\xi - \mu_{\xi,0}\|_{\mathcal{H}_l} = O_p(n^{-\gamma_m} + n^{-1/2}), \qquad \|\widehat{\mu}_X - \mu_{X,0}\|_{\mathcal{H}_k} = O_p(n^{-1/2}).$$

Their product consequently satisfies

$$\|\widehat{\mu}_X - \mu_{X,0}\|_{\mathcal{H}_k}\|\widehat{\mu}_\xi - \mu_{\xi,0}\|_{\mathcal{H}_l} = O_p(n^{-1/2-\gamma_m} + n^{-1}) = o_p(n^{-1/2}).$$

The consistency assumptions and Lemmas 3 and 4 now give, for the nuisance estimates $\widehat{\eta}^{-k}$ fitted outside each evaluation fold I_k,

$$\|D_{\widehat{\eta}^{-k}} - D_{\eta_0}\|_{L^2(P_0;\mathcal{H}_{kl})} = o_p(1), \qquad \|P_0 D_{\widehat{\eta}^{-k}} - \Psi_0\|_{\mathcal{H}_{kl}} = o_p(n^{-1/2}).$$

To obtain the expansion for fixed K, let $n_k = |I_k| = n/K$ and $\mathbb{P}_k = n_k^{-1} \sum_{i \in I_k} \delta_{O_i}$. Conditional on the training observations outside I_k, the evaluation observations are independent. The second moment of their centered average is therefore bounded by

$$\begin{aligned}
&\mathbb{E}_0\left[\|(\mathbb{P}_k - P_0)(D_{\widehat{\eta}^{-k}} - D_{\eta_0})\|^2_{\mathcal{H}_{kl}} \,\middle|\, (O_i)_{i \notin I_k}\right] \\
&\leq \frac{1}{n_k}\|D_{\widehat{\eta}^{-k}} - D_{\eta_0}\|^2_{L^2(P_0;\mathcal{H}_{kl})} = o_p(n^{-1}).
\end{aligned}$$

Conditional Markov's inequality gives $(\mathbb{P}_k - P_0)(D_{\widehat{\eta}^{-k}} - D_{\eta_0}) = o_p(n^{-1/2})$ for each k. Writing $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{O_i}$, we obtain

$$\begin{aligned}
\check{\Psi}_n - \Psi_0 &= \mathbb{P}_n\phi_0 + \frac{1}{K}\sum_{k=1}^{K}(\mathbb{P}_k - P_0)(D_{\widehat{\eta}^{-k}} - D_{\eta_0}) \\
&\quad + \frac{1}{K}\sum_{k=1}^{K}(P_0 D_{\widehat{\eta}^{-k}} - \Psi_0) \\
&= \mathbb{P}_n\phi_0 + o_p(n^{-1/2}).
\end{aligned}$$

The Hilbert-space central limit theorem applies because $\phi_0 \in L^2_0(P_0; \mathcal{H}_{kl})$. □

E.2 A Sobolev example with reused training observations

Corollary 3 (Sobolev sufficient conditions). *Let $\mathcal{W} \subset \mathbb{R}^{d_w}$ be a bounded domain with smooth boundary, where $d_w \geq 1$ is fixed, and suppose $P_{0,W}$ has a density bounded above and below by positive constants on $\mathcal{W}$. Suppose the model, kernel, and conditional moment assumptions in Theorem 3 hold. Use a fixed number $K \geq 2$ of equal evaluation folds, with $N = n(1-1/K)$ training observations for each fit. Assume that there are constants $\sigma_m, R_m > 0$ such that, for every integer $r \geq 2$,*

$$\mathbb{E}_0[\|Y - m_0(W)\|_{\mathbb{R}^{d_y}}^r \mid W] \leq \frac{r!}{2}\sigma_m^2 R_m^{r-2} \quad \text{almost surely.}$$

Let $m_0 \in H^{r_m}(\mathcal{W}; \mathbb{R}^{d_y})$ with $r_m > d_w/2$, and suppose $v_{0,j} \in H^{r_v}(\mathcal{W}; \mathcal{H}_l)$ for every $j \in \{1, \ldots, d_y\}$, where $r_v > 0$. In each outer training fold, fit $\widehat{m}$ by kernel ridge regression with a kernel whose reproducing kernel Hilbert space is norm-equivalent to $H^{a_m}(\mathcal{W})$, where

$$\max\{d_w/2, r_m/2\} < a_m < r_m, \qquad \lambda_{m,n} \asymp n^{-2a_m/(2r_m+d_w)}.$$

Use the same training observations to fit $\widehat{v}_j$ and the residual mean embedding. For $\widehat{v}_j$, use a kernel whose reproducing kernel Hilbert space is norm-equivalent to $H^{a_v}(\mathcal{W})$, where $a_v > d_w/2$ and $0 < r_v \leq 2a_v$, with

$$\lambda_{v,n} \asymp n^{-2a_v/(2r_v+d_w)}.$$

Then

$$\|\widehat{m} - m_0\|_{L^2(P_{0,W};\mathbb{R}^{d_y})} + \|\widehat{m} - m_0\|_{L^2(P^I_{W,N};\mathbb{R}^{d_y})} = O_p(n^{-r_m/(2r_m+d_w)}),$$
$$\|\widehat{m} - m_0\|_{L^\infty(P_{0,W};\mathbb{R}^{d_y})} = o_p(1),$$

and

$$\|\widehat{v}_j - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} = O_p\left(n^{-r_m/(2r_m+d_w)} + n^{-r_v/(2r_v+d_w)}\right). \tag{15}$$

If

$$\frac{r_m}{2r_m + d_w} + \frac{r_v}{2r_v + d_w} > \frac{1}{2}, \tag{16}$$

then $\check{\Psi}_n$ has the asymptotic linear expansion in Corollary 2.

Proof. For $\widehat{m}$, Li et al. [2024, Corollary 2], applied to finite-dimensional outputs with $r_m < 2a_m$, gives

$$\|\widehat{m} - m_0\|_{L^2(P_{0,W};\mathbb{R}^{d_y})} = O_p(n^{-r_m/(2r_m+d_w)}),$$
$$\|\widehat{m} - m_0\|_{H^{a_m}(\mathcal{W};\mathbb{R}^{d_y})} = O_p(n^{-(r_m-a_m)/(2r_m+d_w)}).$$

Since $a_m > d_w/2$, Sobolev embedding yields

$$\|\widehat{m} - m_0\|_{L^\infty(P_{0,W};\mathbb{R}^{d_y})} \lesssim \|\widehat{m} - m_0\|_{H^{a_m}(\mathcal{W};\mathbb{R}^{d_y})}$$
$$= O_p(n^{-(r_m-a_m)/(2r_m+d_w)}) = o_p(1).$$

The exponent is positive because $a_m < r_m$.

We next bound the empirical L^2 error of $\widehat{m}$. The choice of $\lambda_{m,n}$ gives

$$\frac{\log n}{n\lambda_{m,n}} \asymp (\log n) n^{-1+2a_m/(2r_m+d_w)} \longrightarrow 0.$$

Thus Fischer and Steinwart [2020, Lemma 17], applied to the kernel used for $\widehat{m}$ with embedding exponent 1, gives the event (10) for that kernel with probability tending to one. The absolute quadratic-form bound proved in Lemma 5 also gives the reverse operator inequality

$$\widehat{C}_q \preceq \frac{3}{2}C_q + \frac{\lambda}{2}\,\mathrm{Id}_{\mathcal{H}_q}.$$

Applying this inequality to the kernel used for $\widehat{m}$, summing over its d_y coordinates, and using equivalence of the kernel and Sobolev norms yields

$$\begin{aligned}\|\widehat{m} - m_0\|^2_{L^2(P^I_{W,N};\mathbb{R}^{d_y})} &\lesssim \|\widehat{m} - m_0\|^2_{L^2(P_{0,W};\mathbb{R}^{d_y})} \\ &\quad + \lambda_{m,n}\|\widehat{m} - m_0\|^2_{H^{a_m}(\mathcal{W};\mathbb{R}^{d_y})} \\ &= O_p(n^{-2r_m/(2r_m+d_w)}) \\ &\quad + O_p\Big(n^{-2a_m/(2r_m+d_w)} n^{-2(r_m-a_m)/(2r_m+d_w)}\Big) \\ &= O_p(n^{-2r_m/(2r_m+d_w)}).\end{aligned}$$

This application is valid because $m_0 \in H^{r_m} \subset H^{a_m}$ and $\widehat{m}$ belongs to the fitted reproducing kernel Hilbert space.

For $\widetilde{v}_{j,\lambda_{v,n}}$, the regression target is $v_{0,j}$. By Lemma 1 and conditional Jensen's inequality,

$$\|\partial_j \varphi_l\{Y - m_0(W)\}\|_{\mathcal{H}_l} \le \sqrt{2\kappa_l L_2}, \qquad \|v_{0,j}(W)\|_{\mathcal{H}_l} \le \sqrt{2\kappa_l L_2}.$$

Hence, for every integer $r \ge 2$,

$$\mathbb{E}_0\big[\|\partial_j \varphi_l\{Y - m_0(W)\} - v_{0,j}(W)\|^r_{\mathcal{H}_l} \,\big|\, W\big] \le (2\sqrt{2\kappa_l L_2})^r \le \frac{r!}{2}(2\sqrt{2\kappa_l L_2})^r.$$

The conditional Bernstein moment condition therefore holds for these regression labels. Applying Li et al. [2024, Corollary 2] gives

$$\|\widetilde{v}_{j,\lambda_{v,n}} - v_{0,j}\|_{L^2(P_{0,W};\mathcal{H}_l)} = O_p(n^{-r_v/(2r_v+d_w)}).$$

It remains to verify (10) at $\lambda_{v,n}$. By Li et al. [2024, Proposition 4], the Sobolev embedding property holds for any exponent α satisfying $d_w/(2a_v) < \alpha \le 1$. Choose

$$\frac{d_w}{2a_v} < \alpha < \min\left\{1, \frac{2r_v + d_w}{2a_v}\right\}.$$

Such a choice exists because $a_v > d_w/2$ and $r_v > 0$. It gives

$$\frac{\log n}{n\lambda^\alpha_{v,n}} \asymp (\log n) n^{-1+2a_v\alpha/(2r_v+d_w)} \longrightarrow 0.$$

The covariance bound of Fischer and Steinwart [2020, Lemma 17] yields

$$\begin{aligned}&\Big\|(C_q + \lambda_{v,n}\,\mathrm{Id}_{\mathcal{H}_q})^{-1/2}(C_q - \widehat{C}_q)(C_q + \lambda_{v,n}\,\mathrm{Id}_{\mathcal{H}_q})^{-1/2}\Big\|_{\mathrm{op}} \\ &\quad = O_p\left(\sqrt{\frac{\log n}{n\lambda^\alpha_{v,n}}} + \frac{\log n}{n\lambda^\alpha_{v,n}}\right) = o_p(1).\end{aligned}$$

Consequently, (10) holds with probability tending to one. We now apply Corollary 1 to obtain (15). Finally,

$$\frac{r_m}{2r_m + d_w} - \frac{1}{4} = \frac{2r_m - d_w}{4(2r_m + d_w)} > 0,$$

and (16) gives the other inequality in (14). The conclusion follows from Corollary 2. □

F Computation

Notation	Meaning
$\widehat{A}_i, \widehat{C}_i, \widehat{M}_i$	Centered X feature, centered residual feature, and corrected residual feature in the sample-split formulas.
$\widehat{D}_i, G_{ii'}$	Tensor-product one-step summand and its Gram entry.
$\mathcal{D}^k, \mathcal{D}^{-k}, n_k$	Fold k, its complement, and fold size in the cross-fitted computation.
$\mathbb{P}_k, \mathbb{P}_k^{(b)}$	Empirical law on fold k and its bootstrap resample.
$z = (w, y)$	Abbreviation used for the residual-side arguments.
K, $L^{(-k_1,-k_2)}$	Stored Gram matrices for the X kernel and fold-specific residual kernel.
$r_X, r_\xi, r_{\xi,v}, r_v$	Stored centered X term, residual-feature term, residual–v cross term, and v-diagonal term.
∂L, ΔL	Stored first- and mixed-second-derivative tensors for the residual kernel.
Ξ^{-k}, W^{-k}	Stored residual-coordinate tensor and vvKRR weight tensor.
$\mathfrak{w}_j^{-k}$	Scalar vvKRR weight for label coordinate j trained outside fold k.

We first explain the computations in the simpler setting instantiated in Section 3. Then we extend this example to the cross-fitted estimators, where the computations are described in additional detail. Since our experiments focus on the setting where X is W-measurable (Assumption 1), we describe this procedure, to facilitate the reproduction of our results, which use the formulas below. The extension to general W primarily differs in the need to estimate the $\mathcal{H}_{kl}$-valued F_j instead of the $\mathcal{H}_l$-valued v_j, which further leads to slightly different expansions of the inner products analysed below.

F.1 Sample-split computations

For $i \in I_1 \cup I_2$, define $\widehat{\xi}_i \doteq y_i - \widehat{m}(w_i)$. The residuals $\widehat{\xi}_i$ are evaluated for all $i \in I_1 \cup I_2$. The mean embeddings are computed on I_2:

$$\widehat{\mu}_X \doteq \frac{2}{n} \sum_{t \in I_2} \varphi_k(x_t), \qquad \widehat{\mu}_\xi \doteq \frac{2}{n} \sum_{t \in I_2} \varphi_l(\widehat{\xi}_t).$$

Thus $\widehat{\xi}_i$ is used both at evaluation points $i \in I_1$ and at centering points $t \in I_2$. Let $\widehat{v}_j$, $j = 1, \ldots, d_y$, be the remaining nuisance estimates fitted on I_2. For $i \in I_1$, define

$$\widehat{A}_i \doteq \varphi_k(x_i) - \widehat{\mu}_X, \qquad \widehat{C}_i \doteq \varphi_l(\widehat{\xi}_i) - \widehat{\mu}_\xi, \qquad \widehat{M}_i \doteq \widehat{C}_i - \sum_{j=1}^{d_y} \widehat{\xi}_{i,j} \widehat{v}_j(w_i).$$

Then

$$\widehat{D}_i \doteq \widehat{D}(o_i) = \widehat{A}_i \otimes \widehat{M}_i.$$

Hence, the tensor-product inner product factorizes:

$$G_{ii'} \doteq \langle \widehat{D}_i, \widehat{D}_{i'} \rangle_{\mathcal{H}_{kl}} = \langle \widehat{A}_i, \widehat{A}_{i'} \rangle_{\mathcal{H}_k} \langle \widehat{M}_i, \widehat{M}_{i'} \rangle_{\mathcal{H}_l}.$$

Thus $\widehat{Q}_V = \|\widehat{\Psi}_n\|^2_{\mathcal{H}_{kl}} = \frac{4}{n^2} \sum_{i,i' \in I_1} G_{ii'}$ can be computed using only inner products in $\mathcal{H}_k$ and $\mathcal{H}_l$. The residual-side inner product is

$$\begin{aligned} \left\langle \widehat{M}_i, \widehat{M}_{i'} \right\rangle_{\mathcal{H}_l} = \left\langle \widehat{C}_i, \widehat{C}_{i'} \right\rangle_{\mathcal{H}_l} - \sum_{j=1}^{d_y} \widehat{\xi}_{i',j} \left\langle \widehat{C}_i, \widehat{v}_j(w_{i'}) \right\rangle_{\mathcal{H}_l} \\ - \sum_{j=1}^{d_y} \widehat{\xi}_{i,j} \left\langle \widehat{v}_j(w_i), \widehat{C}_{i'} \right\rangle_{\mathcal{H}_l} + \sum_{j,r=1}^{d_y} \widehat{\xi}_{i,j} \widehat{\xi}_{i',r} \left\langle \widehat{v}_j(w_i), \widehat{v}_r(w_{i'}) \right\rangle_{\mathcal{H}_l}. \end{aligned}$$

The first factor is the centered k-Gram entry

$$\left\langle \widehat{A}_i, \widehat{A}_{i'} \right\rangle_{\mathcal{H}_k} = k(x_i, x_{i'}) - \frac{2}{n}\sum_{t\in I_2} k(x_i, x_t) - \frac{2}{n}\sum_{t\in I_2} k(x_{i'}, x_t) + \frac{4}{n^2}\sum_{s,t\in I_2} k(x_s, x_t).$$

Similarly,

$$\left\langle \widehat{C}_i, \widehat{C}_{i'} \right\rangle_{\mathcal{H}_l} = l(\widehat{\xi}_i, \widehat{\xi}_{i'}) - \frac{2}{n}\sum_{t\in I_2} l(\widehat{\xi}_i, \widehat{\xi}_t) - \frac{2}{n}\sum_{t\in I_2} l(\widehat{\xi}_{i'}, \widehat{\xi}_t) + \frac{4}{n^2}\sum_{s,t\in I_2} l(\widehat{\xi}_s, \widehat{\xi}_t).$$

It remains to compute the terms involving $\widehat{v}_j$. Let $R_{j,t} \doteq (\partial_j \varphi_l)(\widehat{\xi}_t), t \in I_2$. The vector-valued KRR estimator has the representer form [Carmeli et al., 2006]

$$\widehat{v}_j(w) = \sum_{t\in I_2} \omega_{j,t}(w) R_{j,t},$$

where, writing $Q = (q(w_s, w_t))_{s,t\in I_2}$ and $q_w = (q(w, w_t))_{t\in I_2}$, $\omega_j(w)^\top = q_w^\top \left(Q + \frac{n}{2}\lambda_{j,n} I\right)^{-1}$. Therefore,

$$\left\langle \widehat{C}_i, \widehat{v}_j(w) \right\rangle_{\mathcal{H}_l} = \sum_{t\in I_2} \omega_{j,t}(w) \left[\partial_{2,j} l(\widehat{\xi}_i, \widehat{\xi}_t) - \frac{2}{n}\sum_{s\in I_2} \partial_{2,j} l(\widehat{\xi}_s, \widehat{\xi}_t)\right],$$

where $\partial_{2,j} l$ denotes the derivative of l with respect to the jth coordinate of its second argument. Finally, $\langle \widehat{v}_j(w), \widehat{v}_r(w') \rangle_{\mathcal{H}_l} = \sum_{s,t\in I_2} \omega_{j,s}(w)\omega_{r,t}(w')\partial_{1,j}\partial_{2,r} l(\widehat{\xi}_s, \widehat{\xi}_t)$. Thus the statistic can be computed from evaluations of k, l, first derivatives $\partial_{2,j} l$, mixed second derivatives $\partial_{1,j}\partial_{2,r} l$, and the scalar KRR weight matrices associated with the kernel q.

F.2 Cross-fitted estimators and confidence sets.

We move on to explain how to compute the point estimates and confidence sets in full generality for K-fold cross-fitted estimators.

Usage of superscripts and subscripts. For an object which depends on estimated nuisances, we always use ***superscripts*** to denote the fold used to fit the nuisances. In addition, we always use ***subscripts*** to denote data points at which an object is evaluated, or indices.

Usage of colour. blue denotes input/data, red denotes intermediate mathematical objects, and green denotes quantities stored for computation. As a rule, we never express objects in red in terms of objects in green. For objects in red, we denote evaluation at data points as function evaluation, namely we use $(\cdot)$ in normal size. In contrast, we view objects in green as tensors, therefore we consider points at which they are evaluated as indices, which are written inside parentheses in subscript. When blue and green clash, green takes precedence.

Data $\mathcal{D} = (X_i, Z_i := (W_i, Y_i))_{i\in[n]}$, split into K folds $\mathcal{D}^k$ of sizes n_k and with complements $\mathcal{D}^{-k}$.

Assumption 5 (Equal fold sizes)**.** *Assume K divides n and all folds are of size $n_k = n/K$.*

Remark 2. *This is assumed purely for simplicity of exposition. It suffices that fold sizes are such that there exist constants $c < C$ for which $c < \frac{n_k}{n} < C$ for all sufficiently large n.*

The cross-fitted estimator of Ψ_0 is

$$\check{\Psi}_n \doteq K^{-1}\sum_{k=1}^{K} \mathbb{P}_k\left(\widehat{D}^{-k}\right).$$

where $\mathbb{P}_k$ is the empirical law of $\mathcal{D}^k$ and $\widehat{D}^{-k}$ is fitted entirely using $\mathcal{D}^{-k}$. Further, we will denote $\mathbb{P}_k^{(b)}$ to be empirical measures obtained by bootstrapping $\mathbb{P}_k$ [Efron, 1992]. We shall denote $\mathbb{P}_k^{(0)} \doteq \mathbb{P}_k$. A minimal objective is to compute quantities of the form $\left\langle \mathbb{P}_{k_1}^{(b_1)}\widehat{D}^{-k_1}, \mathbb{P}_{k_2}^{(b_2)}\widehat{D}^{-k_2} \right\rangle_{\mathcal{H}_{kl}}$, since all

quantities that we need to compute are expressed in terms of inner products of (possibly bootstrapped) estimators of Ψ_0.

We derive the formulas for $b_1 = b_2 = 0$. Write $z = (w, y)$. Recall that for a set of nuisances η,

$$\begin{aligned} A_\eta(x) &\doteq \varphi_k(x) - \mu_X, \\ M_\eta(z) &\doteq \varphi_l(\xi(z)) - \mu_\xi - \sum_{j=1}^{d_y} \xi_j(z) v_j(w), \\ D_\eta(x,z) &= A_\eta(x) \otimes M_\eta(z). \end{aligned}$$

Our estimator should be a plug-in version of these. That is,

$$\widehat{A}^{-k}(x) \doteq \varphi_k(x) - \widehat{\mu}_X^{-k}, \qquad \widehat{M}^{-k}(z) \doteq \varphi_l\left(\widehat{\xi}^{-k}(z)\right) - \widehat{\mu}_\xi^{-k} - \sum_{j=1}^{d_y} \widehat{\xi}_j^{-k}(z) \widehat{v}_j^{-k}(w).$$

$$\widehat{D}^{-k}(x,z) = \widehat{A}^{-k}(x) \otimes \widehat{M}^{-k}(z).$$

Quantities such as

$$\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} = \frac{1}{K^2} \sum_{k_1,k_2} \left\langle \mathbb{P}_{k_1} \widehat{D}^{-k_1}, \mathbb{P}_{k_2} \widehat{D}^{-k_2} \right\rangle_{\mathcal{H}_{kl}}$$

expand as averages of the following V-statistics,

$$\left\langle \mathbb{P}_{k_1} \widehat{D}^{-k_1}, \mathbb{P}_{k_2} \widehat{D}^{-k_2} \right\rangle_{\mathcal{H}_{kl}} = \frac{1}{n_{k_1} n_{k_2}} \sum_{(x_1,z_1) \in \mathcal{D}^{k_1}} \sum_{(x_2,z_2) \in \mathcal{D}^{k_2}} \left\langle \widehat{D}^{-k_1}(x_1,z_1), \widehat{D}^{-k_2}(x_2,z_2) \right\rangle_{\mathcal{H}_{kl}}. \tag{17}$$

We will explain how to compute those summands. Since $\langle f_1 \otimes g_1, f_2 \otimes g_2 \rangle_{\mathcal{H}_{kl}} = \langle f_1, f_2 \rangle_{\mathcal{H}_k} \langle g_1, g_2 \rangle_{\mathcal{H}_l}$ [Aubin, 2000], we isolate two kinds of terms out of these inner products.

F.3 Terms involving covariate features

We define the empirical kernel mean embedding

$$\widehat{\mu}_X^{-k} \doteq \frac{1}{n - n_k} \sum_{x \in \mathcal{X}^{-k}} \varphi_k(x).$$

Thus, for any x_1, x_2, we have

$$\begin{aligned} r^{(-k_1,-k_2)}_{X,(x_1,x_2)} &\doteq \left\langle \varphi_k(x_1) - \widehat{\mu}_X^{-k_1}, \varphi_k(x_2) - \widehat{\mu}_X^{-k_2} \right\rangle_{\mathcal{H}_k} \\ &= k(x_1,x_2) - \frac{1}{n - n_{k_1}} \sum_{x \in \mathcal{X}^{-k_1}} k(x,x_2) - \frac{1}{n - n_{k_2}} \sum_{x \in \mathcal{X}^{-k_2}} k(x,x_1) \\ &\quad + \frac{1}{(n - n_{k_1})(n - n_{k_2})} \sum_{\substack{x \in \mathcal{X}^{-k_1} \\ x' \in \mathcal{X}^{-k_2}}} k(x,x') \\ &= K_{(x_1,x_2)} - \frac{1}{n - n_{k_1}} 1_{n-n_{k_1}}^\top K_{(-k_1,x_2)} - \frac{1}{n - n_{k_2}} K_{(x_1,-k_2)} 1_{n-n_{k_2}} \\ &\quad + \frac{1}{(n - n_{k_1})(n - n_{k_2})} 1_{n-n_{k_1}}^\top K_{(-k_1,-k_2)} 1_{n-n_{k_2}}, \end{aligned}$$

where $K \in \mathbb{R}^{n \times n}$ denotes the full-data kernel matrix on X, namely $K_{(x_1,x_2)} = k(x_1,x_2)$, and 1_d denotes the $d \times 1$ vector of all ones, for any $d \geq 1$. $K_{(-k_1,x_2)} \in \mathbb{R}^{(n-n_{k_1}) \times 1}$ denotes the matrix of $k(x,x_2)$, where x are the X observations in fold $\mathcal{X}^{-k_1}$. The remaining notations are defined analogously.

F.4 Terms involving residual features

We continue to write $z = (w, y)$, e.g. $z_1 = (w_1, y_1)$.

Analogously to the X-setting, we define

$$\widehat{\mu}_\xi^{-k} \doteq \frac{1}{n - n_k} \sum_{z \in \mathcal{Z}^{-k}} \varphi_l \left\{ \widehat{\xi}^{-k}(z) \right\}.$$

For any z, we partition $\widehat{M}^{-k}(z) \in \mathcal{H}_l$ into two summands,

$$\varphi_l(\widehat{\xi}^{-k}(z)) - \widehat{\mu}_\xi^{-k}, \qquad -\sum_{j=1}^{d_y} \underbrace{\widehat{\xi}_j^{-k}(z)}_{\mathbb{R}} \underbrace{\widehat{v}_j^{-k}(w)}_{\mathcal{H}_l}.$$

That yields two kinds of "diagonal terms" and one type of "cross-term" in the inner product decomposition

$$\left\langle \widehat{M}^{-k_1}(z_1), \widehat{M}^{-k_2}(z_2) \right\rangle_{\mathcal{H}_l}.$$

φ_l-diagonal terms. Let $L^{(-k_1,-k_2)} \in \mathbb{R}^{n \times n}$ have entries

$$L^{(-k_1,-k_2)}_{(z_1,z_2)} \doteq \left\langle \varphi_l \left(\widehat{\xi}^{-k_1}(z_1) \right), \varphi_l \left(\widehat{\xi}^{-k_2}(z_2) \right) \right\rangle_{\mathcal{H}_l} = l \left(\widehat{\xi}^{-k_1}(z_1), \widehat{\xi}^{-k_2}(z_2) \right).$$

We expand the inner product of residual features as

$$\begin{aligned} r^{(-k_1,-k_2)}_{\xi,(z_1,z_2)} &\doteq \left\langle \varphi_l \left(\widehat{\xi}^{-k_1}(z_1) \right) - \widehat{\mu}_\xi^{-k_1}, \varphi_l \left(\widehat{\xi}^{-k_2}(z_2) \right) - \widehat{\mu}_\xi^{-k_2} \right\rangle_{\mathcal{H}_l} \\ &= L^{(-k_1,-k_2)}_{(z_1,z_2)} - \frac{1}{n - n_{k_1}} \mathbf{1}_{n-n_{k_1}}^\intercal L^{(-k_1,-k_2)}_{(-k_1,z_2)} - \frac{1}{n - n_{k_2}} L^{(-k_1,-k_2)}_{(z_1,-k_2)} \mathbf{1}_{n-n_{k_2}} \\ &+ \frac{1}{(n - n_{k_1})(n - n_{k_2})} \mathbf{1}_{n-n_{k_1}}^\intercal L^{(-k_1,-k_2)}_{(-k_1,-k_2)} \mathbf{1}_{n-n_{k_2}}, \end{aligned}$$

where $r_\xi^{(-k_1,-k_2)} \in \mathbb{R}^{n_{k_1} \times n_{k_2}}$.

Cross-terms. We define

$$r^{(-k_1,-k_2)}_{\xi,v,(z_1,z_2)} \doteq \left\langle \varphi_l \left(\widehat{\xi}^{-k_1}(z_1) \right) - \widehat{\mu}_\xi^{-k_1}, \sum_{j=1}^{d_y} \widehat{\xi}_j^{-k_2}(z_2) \widehat{v}_j^{-k_2}(w_2) \right\rangle_{\mathcal{H}_l}, \tag{18}$$

and an analogous term with $1 \leftrightarrow 2$. Suppose that $\widehat{v}_j^{-k}$ is the solution to a vector-valued kernel ridge regression with separable kernel [Li et al., 2024]

$$Q(w_1, w_2) = q(w_1, w_2) \mathrm{Id}_{\mathcal{H}_l}.$$

Then it can be written as

$$\widehat{v}_j^{-k}(w) = \sum_{z' \in \mathcal{Z}^{-k}} \mathfrak{w}_j^{-k}(z', w) (\partial_j \varphi_l) \left(\widehat{\xi}^{-k}(z') \right), \tag{19}$$

where $z' = (w', y')$ and

$$\mathfrak{w}_j^{-k}(z', w) = \left[\left(Q^{-k} + (n - n_k) \lambda_{j,n} I \right)^{-1} q^{-k}(w) \right]_{z'},$$

with

$$(Q^{-k})_{z',z''} = q(w', w''), \qquad (q^{-k}(w))_{z''} = q(w'', w).$$

Remark 3. *The vvKRR fit above is only one way to estimate v_j^{-k}. Any regression method can be used provided that the inner products appearing below can be computed. A particularly useful variant is to center the regression targets. Let*

$$S_j^{-k}(z') := (\partial_j \varphi_l)\left(\widehat{\xi}^{-k}(z')\right), \qquad \bar{S}_j^{-k} := \frac{1}{n-n_k} \sum_{z' \in \mathcal{Z}^{-k}} S_j^{-k}(z').$$

If vvKRR is fit to the centered labels $S_j^{-k}(z') - \bar{S}_j^{-k}$, we add the mean back:

$$\widehat{v}_j^{-k}(w) = \bar{S}_j^{-k} + \sum_{z' \in \mathcal{Z}^{-k}} \widetilde{\mathfrak{w}}_j^{-k}(z', w)\{S_j^{-k}(z') - \bar{S}_j^{-k}\}.$$

Equivalently,

$$\widehat{v}_j^{-k}(w) = \sum_{z' \in \mathcal{Z}^{-k}} \mathfrak{w}_{j,\mathrm{cen}}^{-k}(z', w) S_j^{-k}(z'),$$

where

$$\mathfrak{w}_{j,\mathrm{cen}}^{-k}(z', w) = \widetilde{\mathfrak{w}}_j^{-k}(z', w) + \frac{1 - \sum_{u \in \mathcal{Z}^{-k}} \widetilde{\mathfrak{w}}_j^{-k}(u, w)}{n - n_k}.$$

Thus the formulas below are unchanged after replacing $\mathfrak{w}_j^{-k}$ by $\mathfrak{w}_{j,\mathrm{cen}}^{-k}$. Centering is particularly helpful under the $\Psi_0 = 0$ null in a setting where $W = X$. There $v_{0,j}$ is constant. The constant component is then estimated by the empirical mean and is not penalized by the ridge regression.

We define

$$\partial L_{j,(z_1,z_2)}^{(-k_1,-k_2)} = \left(\partial_j^1 l\right)\left(\widehat{\xi}^{-k_1}(z_1), \widehat{\xi}^{-k_2}(z_2)\right),$$

where $\partial L^{(-k_1,-k_2)} \in \mathbb{R}^{d_y \times n \times n}$ is a tensor of rank 3, and ∂_j^1 denotes the partial derivative with respect to the j-th component of the first $\mathbb{R}^{d_y}$-valued argument. It is not in general symmetric in its second and third coordinates. We also emphasize that ∂L should be understood as a single symbol. Substituting Eq. (19) into Eq. (18) yields

$$\begin{aligned}
&r_{\xi,v,(z_1,z_2)}^{(-k_1,-k_2)} \\
&= \left\langle \varphi_l\left(\widehat{\xi}^{-k_1}(z_1)\right) - \widehat{\mu}_\xi^{-k_1}, \sum_{j=1}^{d_y} \sum_{z' \in \mathcal{Z}^{-k_2}} \mathfrak{w}_j^{-k_2}(z', w_2)(\partial_j \varphi_l)\left(\widehat{\xi}^{-k_2}(z')\right) \widehat{\xi}_j^{-k_2}(z_2) \right\rangle_{\mathcal{H}_l} \\
&= \sum_{j=1}^{d_y} \sum_{z' \in \mathcal{Z}^{-k_2}} \mathfrak{w}_j^{-k_2}(z', w_2) \widehat{\xi}_j^{-k_2}(z_2) \left\langle (\partial_j \varphi_l)\left(\widehat{\xi}^{-k_2}(z')\right), \varphi_l\left(\widehat{\xi}^{-k_1}(z_1)\right) - \widehat{\mu}_\xi^{-k_1} \right\rangle_{\mathcal{H}_l} \\
&= \sum_{j=1}^{d_y} \sum_{z' \in \mathcal{Z}^{-k_2}} \mathfrak{w}_j^{-k_2}(z', w_2) \widehat{\xi}_j^{-k_2}(z_2) \left\{ \partial L_{j,(z',z_1)}^{(-k_2,-k_1)} - \frac{1}{n-n_{k_1}} \partial L_{j,(z',-k_1)}^{(-k_2,-k_1)\top} 1_{n-n_{k_1}} \right\}.
\end{aligned}$$

Next, we define tensors storing the quantities needed to compute $\mathfrak{w}_j^{-k_2}(z', w_2)\widehat{\xi}_j^{-k_2}(z_2)$. Define a rank 3-tensor $W^{-k} \in \mathbb{R}^{d_y \times (n-n_k) \times n}$ and rank 2-tensor $\Xi^{-k} \in \mathbb{R}^{d_y \times n}$ by

$$\begin{aligned}
\Xi_{j,z}^{-k} &\doteq \widehat{\xi}_j^{-k}(z) \\
W_{j,(z',w)}^{-k} &\doteq \mathfrak{w}_j^{-k}(z', w).
\end{aligned}$$

We can therefore write

$$\begin{aligned}
&r_{\xi,v,(z_1,z_2)}^{(-k_1,-k_2)} \\
&= \sum_{j=1}^{d_y} \sum_{z' \in \mathcal{Z}^{-k_2}} W_{j,(z',w_2)}^{-k_2} \Xi_{j,z_2}^{-k_2} \left\{ \partial L_{j,(z',z_1)}^{(-k_2,-k_1)} - \frac{1}{n-n_{k_1}} \partial L_{j,(z',-k_1)}^{(-k_2,-k_1)\top} 1_{n-n_{k_1}} \right\}.
\end{aligned}$$

The above tensor contraction can be performed using one line of Einstein summation in NumPy or JAX.

v-diagonal terms. We define the rank-4 tensor $\Delta L^{(-k_1,-k_2)} \in \mathbb{R}^{d_y \times d_y \times (n-n_{k_1}) \times (n-n_{k_2})}$ as follows

$$\Delta L^{(-k_1,-k_2)}_{j_1,j_2,(z',z'')} = \partial^1_{j_1} \partial^2_{j_2} l \left(\widehat{\xi}^{-k_1}(z'), \widehat{\xi}^{-k_2}(z'') \right).$$

Therefore we have

$$\begin{aligned}
& r^{(-k_1,-k_2)}_{v,(z_1,z_2)} \\
&= \left\langle \sum_{j_1=1}^{d_y} \widehat{\xi}^{-k_1}_{j_1}(z_1) \widehat{v}^{-k_1}_{j_1}(w_1), \sum_{j_2=1}^{d_y} \widehat{\xi}^{-k_2}_{j_2}(z_2) \widehat{v}^{-k_2}_{j_2}(w_2) \right\rangle_{\mathcal{H}_l} \\
&= \sum_{j_1=1}^{d_y} \sum_{j_2=1}^{d_y} \Xi^{-k_1}_{j_1,z_1} \Xi^{-k_2}_{j_2,z_2} \sum_{\substack{z' \in \mathcal{Z}^{-k_1} \\ z'' \in \mathcal{Z}^{-k_2}}} W^{-k_1}_{j_1,(z',w_1)} W^{-k_2}_{j_2,(z'',w_2)} \Delta L^{(-k_1,-k_2)}_{j_1,j_2,(z',z'')},
\end{aligned}$$

which can again be computed using Einstein summation.

For consistency with Section 3, define the fold-pair Gram entry

$$G^{(-k_1,-k_2)}_{((x_1,z_1),(x_2,z_2))} \doteq r^{(-k_1,-k_2)}_{X,(x_1,x_2)} \left\{ r^{(-k_1,-k_2)}_{\xi,(z_1,z_2)} - r^{(-k_1,-k_2)}_{\xi,v,(z_1,z_2)} - r^{(-k_2,-k_1)}_{\xi,v,(z_2,z_1)} + r^{(-k_1,-k_2)}_{v,(z_1,z_2)} \right\}.$$

Putting everything together, we have

$$\left\langle \mathbb{P}_{k_1} \widehat{D}^{-k_1}, \mathbb{P}_{k_2} \widehat{D}^{-k_2} \right\rangle_{\mathcal{H}_{kl}} \tag{20}$$

$$= \frac{1}{n_{k_1} n_{k_2}} \sum_{\substack{(x_1,z_1) \in \mathcal{D}^{k_1} \\ (x_2,z_2) \in \mathcal{D}^{k_2}}} G^{(-k_1,-k_2)}_{((x_1,z_1),(x_2,z_2))}. \tag{21}$$

G Inference

Notation	Meaning
$\mathfrak{O}$, Ω_n, Ω_0	Class of continuous, positive-definite self-adjoint linear operators from $\mathcal{H}_{kl}$ to $\mathcal{H}_{kl}$; empirical and population choices for the operator used to construct a confidence set.
$w(h;\Omega)$, $\mathcal{C}_n(\zeta)$	Quadratic form and resulting Hilbert-space confidence set.
$\mathbb{H}$, $\mathbb{H}_n^{(b)}$	Gaussian limit and bootstrap empirical-process draw.
$\zeta_{1-\alpha}$, $\widehat{\zeta}_n$	Population and bootstrap-estimated quantiles for the chosen confidence-set geometry.
σ^2, $\widehat{\sigma}^2$	Delta-method variance component and its plug-in estimator.
$\check{\Psi}_n$	Cross-fitted one-step estimator of Ψ_0.
$\widehat{D}^{-k}$	Uncentered EIF estimate fitted without fold k.
D_{η_0}, ϕ_0	Population uncentered and centered EIFs, with $\phi_0 = D_{\eta_0} - \Psi_0$.
G_{ij}	Cross-fitted Gram entry $\langle \widehat{D}^{-k(i)}(O_i), \widehat{D}^{-k(j)}(O_j) \rangle_{\mathcal{H}_{kl}}$, where $k(i)$ is the validation fold containing O_i.
$\widehat{Q}_V$, $\widehat{Q}_U$	V-statistic and U-statistic for $\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}}$ and its diagonal-term-free analogue.

We build on the existing inferential framework in Luedtke and Chung [2024, Section 4.2] for Hilbert-valued one-step estimators in order to construct an $(1-\alpha)$-Hilbert-valued confidence set centered at the one-step estimator. Our presentation here is mainly for completeness and ease of reference. We refer the interested reader to the discussion on U-statistic and V-statistic towards the end of this section, which we believe has not been observed before in the literature.

We first import the general machinery from Luedtke and Chung [2024], which illustrates the flexibility with which we can construct our confidence set, before specializing to the construction given in the main text. We let $\mathfrak{O}$ denote the set of continuous, self-adjoint, positive definite linear operators mapping from $\mathcal{H}_{kl}$ to $\mathcal{H}_{kl}$. For $\Omega \in \mathfrak{O}$, we define the quadratic form w acting on $h \in \mathcal{H}_{kl}$ as

$$w(h; \Omega) := \langle h, \Omega h \rangle_{\mathcal{H}_{kl}}.$$

Let $\zeta \geq 0$ be a specified threshold and $\Omega_n \in \mathfrak{O}$ be an estimator of an operator $\Omega_0 \in \mathfrak{O}$, where the subscript 0 reflects its (possible) dependence on the distribution P_0. We define the confidence set

$$\mathcal{C}_n(\zeta) := \left\{ h \in \mathcal{H}_{kl} : w(\check{\Psi}_n - h; \Omega_n) \leq \frac{\zeta}{n} \right\}.$$

We then obtain an estimator $\widehat{\zeta}_n$ of the $(1-\alpha)$-quantile $\zeta_{1-\alpha}$ of $w(\mathbb{H}; \Omega_0)$ via Efron's bootstrap [Efron, 1992]. We show below in Proposition 2 that $\widehat{\zeta}_n$ converges in probability to $\zeta_{1-\alpha}$, where we directly adapt Luedtke and Chung [2024, Theorem 4]. This result naturally and crucially relies on the existence of the limit distribution $\mathbb{H}$ in the RKHS $\mathcal{H}_{kl}$. We then invoke Luedtke and Chung [2024, Theorem 3], which shows that $\mathcal{C}_n(\widehat{\zeta}_n)$ has asymptotic coverage of $1-\alpha$, namely the nominal coverage.

We now adapt the description of Efron's bootstrap in Luedtke and Chung [2024] to our notation, for ease of reference for the reader. We work in the general K-fold cross fitting setting, for $K \geq 2$. For fold $k \in [K]$, we define the empirical measure of the data points lying in fold k with n_k data points as

$$\mathbb{P}_k \doteq n_k^{-1} \sum_{i:k(i)=k} \delta_{O_i},$$

where $k(i)$ selects the fold index for observation i. For simplicity, we assume all folds have exactly the same size $n_k = \frac{n}{K}$. We draw B Monte Carlo samples from the empirical law $\mathbb{P}_k$ as follows. For $b \in [B]$ and $k \in [K]$, we resample with repetition from the empirical measure $\mathbb{P}_k$

$$O_{1,k}^{(b)}, \ldots, O_{n_k,k}^{(b)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_k,$$

and we define the above bootstrap replicate's empirical measure as

$$\mathbb{P}_k^{(b)} \doteq \frac{1}{n_k} \sum_{i=1}^{n_k} \delta_{O_{i,k}^{(b)}}.$$

For $b \in [B]$, we define

$$\mathbb{H}_n^{(b)} := \sqrt{n} \frac{1}{K} \sum_{k=1}^{K} (\mathbb{P}_k^{(b)} - \mathbb{P}_k) \widehat{D}^{-k} \in \mathcal{H}_{kl}.$$

We define $\widehat{\zeta}_n$ as the empirical $(1-\alpha)$-quantile of $w(\mathbb{H}_n^{(b)}; \Omega_n)$ over $b = 1, \ldots, B$. The empirical quantile $\widehat{\zeta}_n$ is an approximation of the $(1-\alpha)$-quantile of the random variable $w(\mathbb{H}_n^{(b)}; \Omega_n)$ (for any b) conditioned on the samples in fold k. We note as in Luedtke and Chung [2024] that an attractive property of this procedure is that $\widehat{D}^{-k}$ is not being refitted B times, which avoids potentially computationally expensive refitting of nonparametric nuisance functions for a large number of bootstrap replicates. The following proposition is a direct application of Luedtke and Chung [2024, Theorem 4].

Proposition 2. *Suppose the conditions of Theorem 3 hold. Further suppose that* $\Omega_n, \Omega_0 \in \mathfrak{O}$, $\|\Omega_n - \Omega_0\|_{\text{op}} = o_p(1)$, *and*

$$\max_{k \in [K]} \left\| \widehat{D}^{-k} - D_{\eta_0} \right\|_{L^2(P_0; \mathcal{H}_{kl})} = o_p(1).$$

We assume that $B \to \infty$. *Then* $\widehat{\zeta}_n \xrightarrow{p} \zeta_{1-\alpha}$.

In this paper, we focus on the case $\Omega_n = \Omega_0 = \text{Id}_{\mathcal{H}_{kl}}$, for which

$$w(\mathbb{H}_n^{(b)}; \Omega_n) = \|\mathbb{H}_n^{(b)}\|_{\mathcal{H}_{kl}}^2.$$

We recover classical HSIC by taking the squared $\mathcal{H}_{kl}$-norm. The squared norm is computed using the V-statistic in Appendix F. A direction which is interesting for future investigation is to study whether we can improve the power of our test by carefully designing Ω_n, possibly in a data-driven way. A particularly natural choice is the regularized Wald-type confidence set described in Luedtke and Chung [2024].

Confidence intervals based on triangle inequality. Let $k(i)$ denote the validation fold containing O_i, and define

$$G_{ij} := \left\langle \widehat{D}^{-k(i)}(O_i), \widehat{D}^{-k(j)}(O_j) \right\rangle_{\mathcal{H}_{kl}}.$$

Then the plug-in squared norm of the cross-fitted estimator is the V-statistic

$$\widehat{Q}_V := \|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} = \frac{1}{n^2} \sum_{i,j=1}^{n} G_{ij}. \tag{22}$$

The quantities G_{ij} are precisely the foldwise Gram entries computed in Appendix F.

If $N_i^{(b)}$ denotes the number of times O_i appears in the bootstrap draw of its validation fold, then, for $\Omega_n = \mathrm{Id}_{\mathcal{H}_{kl}}$,

$$\|\mathbb{H}_n^{(b)}\|^2_{\mathcal{H}_{kl}} = \frac{1}{n} \sum_{i,j=1}^{n} \{N_i^{(b)} - 1\}\{N_j^{(b)} - 1\} G_{ij}.$$

For $\Omega_n = \mathrm{Id}_{\mathcal{H}_{kl}}$, the confidence set is

$$\mathcal{C}_n(\zeta) = \left\{ h \in \mathcal{H}_{kl} : \|\check{\Psi}_n - h\|^2_{\mathcal{H}_{kl}} \le \frac{\zeta}{n} \right\}.$$

Hence, by the reverse triangle inequality, it induces the conservative interval

$$\mathcal{I}_\zeta := \left[\left\{ \sqrt{\widehat{Q}_V} - \sqrt{\zeta/n} \right\}_+, \sqrt{\widehat{Q}_V} + \sqrt{\zeta/n} \right].$$

In implementation we take $\zeta = \widehat{\zeta}_n$. The induced test of $H_0 : \Psi_0 = 0$ rejects when $0 \notin \mathcal{I}_{\widehat{\zeta}_n}$, equivalently when $n\widehat{Q}_V > \widehat{\zeta}_n$.

G.1 Confidence intervals based on the delta method

Beyond testing $\Psi_0 = 0$, we may be interested in the magnitude of the signal, namely $\|\Psi_0\|_{\mathcal{H}_{kl}}$ or $\|\Psi_0\|^2_{\mathcal{H}_{kl}}$. When $\Psi_0 \neq 0$, the functional delta method [van der Vaart, 2000, Chapter 3] gives an asymptotic normal approximation. Under the assumptions of Theorem 3,

$$\sqrt{n}(\check{\Psi}_n - \Psi_0) \rightsquigarrow \mathbb{H}.$$

Since $h \mapsto \|h\|^2_{\mathcal{H}_{kl}}$ is Fréchet differentiable with derivative $u \mapsto 2\langle h, u\rangle_{\mathcal{H}_{kl}}$, we have

$$\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right) \rightsquigarrow 2\langle \Psi_0, \mathbb{H}\rangle_{\mathcal{H}_{kl}}.$$

The limiting variance is $4\sigma^2$, where

$$\sigma^2 := \mathbb{E}_0\left[\langle \Psi_0, \phi_0(O)\rangle^2_{\mathcal{H}_{kl}}\right] = \mathbb{E}_0\left[\langle \Psi_0, D_{\eta_0}(O)\rangle^2_{\mathcal{H}_{kl}}\right] - \|\Psi_0\|^4_{\mathcal{H}_{kl}}.$$

This motivates the estimator

$$\begin{aligned}
\widehat{\sigma}^2 &:= \frac{1}{K}\sum_{l=1}^{K} \frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \left\langle \check{\Psi}_n, \widehat{D}^{-l}(o') \right\rangle^2_{\mathcal{H}_{kl}} - \|\check{\Psi}_n\|^4_{\mathcal{H}_{kl}} \\
&= \frac{1}{K}\sum_{l=1}^{K} \frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \left(\frac{1}{K}\sum_{k=1}^{K} \frac{1}{n_k} \sum_{o \in \mathcal{D}^k} \left\langle \widehat{D}^{-k}(o), \widehat{D}^{-l}(o') \right\rangle_{\mathcal{H}_{kl}} \right)^2 \\
&\quad - \left(\frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K} \frac{1}{n_k n_l} \sum_{\substack{o \in \mathcal{D}^k \\ o' \in \mathcal{D}^l}} \left\langle \widehat{D}^{-k}(o), \widehat{D}^{-l}(o') \right\rangle_{\mathcal{H}_{kl}} \right)^2.
\end{aligned} \tag{23}$$

The inner products $\langle \widehat{D}^{-k}(o), \widehat{D}^{-l}(o') \rangle_{\mathcal{H}_{kl}}$ are precisely the quantities computed in Appendix F.

Proposition 3 (Consistency of the estimator of asymptotic variance)**.** *Suppose the assumptions of Theorem 3 hold, and assume K is fixed and $n_k = n/K$ for all $k \in [K]$. Suppose further that*

$$\max_{k \in [K]} \|\widehat{D}^{-k} - D_{\eta_0}\|_{L^2(P_0; \mathcal{H}_{kl})} = o_p(1),$$

and that

$$\max_{k \in [K]} \int \|\widehat{D}^{-k}(o)\|_{\mathcal{H}_{kl}}^4 \, dP_0(o) = O_p(1), \qquad \int \|D_{\eta_0}(o)\|_{\mathcal{H}_{kl}}^4 \, dP_0(o) < \infty.$$

Define

$$\widehat{\sigma}^2 := \frac{1}{K} \sum_{l=1}^{K} \frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \langle \check{\Psi}_n, \widehat{D}^{-l}(o') \rangle_{\mathcal{H}_{kl}}^2 - \|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^4.$$

Then

$$\widehat{\sigma}^2 \xrightarrow{p} \sigma^2,$$

where

$$\sigma^2 := \mathbb{E}_0[\langle \Psi_0, \phi_0(O) \rangle_{\mathcal{H}_{kl}}^2] = \mathbb{E}_0[\langle \Psi_0, D_{\eta_0}(O) \rangle_{\mathcal{H}_{kl}}^2] - \|\Psi_0\|_{\mathcal{H}_{kl}}^4.$$

Proof. By Theorem 3, $\check{\Psi}_n \xrightarrow{p} \Psi_0$. Hence, by the continuous mapping theorem,

$$\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^4 \xrightarrow{p} \|\Psi_0\|_{\mathcal{H}_{kl}}^4.$$

It remains to show that

$$\frac{1}{K} \sum_{l=1}^{K} \frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \langle \check{\Psi}_n, \widehat{D}^{-l}(o') \rangle_{\mathcal{H}_{kl}}^2 \xrightarrow{p} \mathbb{E}_0[\langle \Psi_0, D_{\eta_0}(O) \rangle_{\mathcal{H}_{kl}}^2].$$

We consider the decomposition

$$\begin{aligned}
&\frac{1}{K} \sum_{l=1}^{K} \frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \langle \check{\Psi}_n, \widehat{D}^{-l}(o') \rangle_{\mathcal{H}_{kl}}^2 - \mathbb{E}_0[\langle \Psi_0, D_{\eta_0}(O) \rangle_{\mathcal{H}_{kl}}^2] \\
&= \frac{1}{K} \sum_{l=1}^{K} \frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \left\{ \langle \check{\Psi}_n, \widehat{D}^{-l}(o') \rangle_{\mathcal{H}_{kl}}^2 - \langle \Psi_0, \widehat{D}^{-l}(o') \rangle_{\mathcal{H}_{kl}}^2 \right\} \qquad \text{(A)} \\
&\quad + \frac{1}{K} \sum_{l=1}^{K} \frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \left\{ \langle \Psi_0, \widehat{D}^{-l}(o') \rangle_{\mathcal{H}_{kl}}^2 - \langle \Psi_0, D_{\eta_0}(o') \rangle_{\mathcal{H}_{kl}}^2 \right\} \qquad \text{(B)} \\
&\quad + \frac{1}{K} \sum_{l=1}^{K} \frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \langle \Psi_0, D_{\eta_0}(o') \rangle_{\mathcal{H}_{kl}}^2 - \mathbb{E}_0[\langle \Psi_0, D_{\eta_0}(O) \rangle_{\mathcal{H}_{kl}}^2]. \qquad \text{(C)}
\end{aligned}$$

For term (A), we use $a^2 - b^2 = (a-b)(a+b)$. We have

$$\begin{aligned}
|\text{(A)}| &\leq \frac{1}{K} \sum_{l=1}^{K} \frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \left| \langle \check{\Psi}_n - \Psi_0, \widehat{D}^{-l}(o') \rangle_{\mathcal{H}_{kl}} \right| \left| \langle \check{\Psi}_n + \Psi_0, \widehat{D}^{-l}(o') \rangle_{\mathcal{H}_{kl}} \right| \\
&\leq \|\check{\Psi}_n - \Psi_0\|_{\mathcal{H}_{kl}} \|\check{\Psi}_n + \Psi_0\|_{\mathcal{H}_{kl}} \frac{1}{K} \sum_{l=1}^{K} \frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \|\widehat{D}^{-l}(o')\|_{\mathcal{H}_{kl}}^2.
\end{aligned}$$

We now show that the final factor is $O_p(1)$. Conditionally on the training sample $\mathcal{D}^{-l}$, the observations in $\mathcal{D}^l$ are i.i.d. under P_0 and independent of $\widehat{D}^{-l}$. Therefore, by Chebyshev's inequality,

$$\frac{1}{n_l} \sum_{o' \in \mathcal{D}^l} \|\widehat{D}^{-l}(o')\|_{\mathcal{H}_{kl}}^2 - \int \|\widehat{D}^{-l}(o)\|_{\mathcal{H}_{kl}}^2 \, dP_0(o) = o_p(1),$$

because

$$\int \|\widehat{D}^{-l}(o)\|_{\mathcal{H}_{kl}}^4 \, dP_0(o) = O_p(1),$$

and $n_l \to \infty$. Since K is fixed, averaging over l gives

$$\frac{1}{K}\sum_{l=1}^{K}\left\{\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o')\|_{\mathcal{H}_{kl}}^2 - \int \|\widehat{D}^{-l}(o)\|_{\mathcal{H}_{kl}}^2 \, dP_0(o)\right\} = o_p(1).$$

On the other hand,

$$\begin{aligned}\frac{1}{K}\sum_{l=1}^{K}\int \|\widehat{D}^{-l}(o)\|_{\mathcal{H}_{kl}}^2 \, dP_0(o) &\le \frac{2}{K}\sum_{l=1}^{K}\|\widehat{D}^{-l} - D_{\eta_0}\|_{L^2(P_0;\mathcal{H}_{kl})}^2 + 2\|D_{\eta_0}\|_{L^2(P_0;\mathcal{H}_{kl})}^2 \\ &= O_p(1).\end{aligned}$$

Therefore

$$\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o')\|_{\mathcal{H}_{kl}}^2 = O_p(1).$$

Since $\check{\Psi}_n \xrightarrow{p} \Psi_0$, we have

$$|(\text{A})| = o_p(1).$$

For term (B), we again use $a^2 - b^2 = (a-b)(a+b)$. We have

$$\begin{aligned}|(\text{B})| &\le \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\Psi_0\|_{\mathcal{H}_{kl}}^2\|\widehat{D}^{-l}(o') - D_{\eta_0}(o')\|_{\mathcal{H}_{kl}}\|\widehat{D}^{-l}(o') + D_{\eta_0}(o')\|_{\mathcal{H}_{kl}} \\ &\le \|\Psi_0\|_{\mathcal{H}_{kl}}^2\left\{\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o') - D_{\eta_0}(o')\|_{\mathcal{H}_{kl}}^2\right\}^{1/2} \\ &\quad\times\left\{\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o') + D_{\eta_0}(o')\|_{\mathcal{H}_{kl}}^2\right\}^{1/2}.\end{aligned}$$

Conditionally on $\mathcal{D}^{-l}$, another application of Chebyshev's inequality gives

$$\frac{1}{K}\sum_{l=1}^{K}\left\{\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o') - D_{\eta_0}(o')\|_{\mathcal{H}_{kl}}^2 - \int \|\widehat{D}^{-l}(o) - D_{\eta_0}(o)\|_{\mathcal{H}_{kl}}^2 \, dP_0(o)\right\} = o_p(1),$$

and similarly with $+$ in place of $-$. Hence

$$\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o') - D_{\eta_0}(o')\|_{\mathcal{H}_{kl}}^2 = \frac{1}{K}\sum_{l=1}^{K}\|\widehat{D}^{-l} - D_{\eta_0}\|_{L^2(P_0;\mathcal{H}_{kl})}^2 + o_p(1) = o_p(1),$$

while

$$\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\|\widehat{D}^{-l}(o') + D_{\eta_0}(o')\|_{\mathcal{H}_{kl}}^2 = O_p(1).$$

Therefore

$$|(\text{B})| = o_p(1).$$

Finally, for term (C), since $D_{\eta_0} \in L^2(P_0;\mathcal{H}_{kl})$, we have

$$\mathbb{E}_0[\langle\Psi_0, D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2] \le \|\Psi_0\|_{\mathcal{H}_{kl}}^2\mathbb{E}_0[\|D_{\eta_0}(O)\|_{\mathcal{H}_{kl}}^2] < \infty.$$

Therefore, by the weak law of large numbers,

$$(\text{C}) = \mathbb{P}_n\{\langle\Psi_0, D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2\} - \mathbb{E}_0[\langle\Psi_0, D_{\eta_0}(O)\rangle_{\mathcal{H}_{kl}}^2] = o_p(1).$$

Combining the bounds for (A), (B), and (C), we obtain

$$\frac{1}{K}\sum_{l=1}^{K}\frac{1}{n_l}\sum_{o'\in\mathcal{D}^l}\langle\check{\Psi}_n,\widehat{D}^{-l}(o')\rangle^2_{\mathcal{H}_{kl}}\xrightarrow{p}\mathbb{E}_0[\langle\Psi_0,D_{\eta_0}(O)\rangle^2_{\mathcal{H}_{kl}}].$$

Together with $\|\check{\Psi}_n\|^4_{\mathcal{H}_{kl}}\xrightarrow{p}\|\Psi_0\|^4_{\mathcal{H}_{kl}}$, this proves

$$\widehat{\sigma}^2\xrightarrow{p}\mathbb{E}_0[\langle\Psi_0,D_{\eta_0}(O)\rangle^2_{\mathcal{H}_{kl}}]-\|\Psi_0\|^4_{\mathcal{H}_{kl}}=\sigma^2.$$

□

Theorem 4 (Validity and efficiency of the delta method)**.** *Suppose the assumptions of Theorem 3 hold. Then, in $\mathcal{H}_{kl}$,*

$$\sqrt{n}(\check{\Psi}_n-\Psi_0)\rightsquigarrow\mathbb{H},$$

where $\mathbb{H}$ is the centered Gaussian element in Theorem 3. Consequently,

$$\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}}-\|\Psi_0\|^2_{\mathcal{H}_{kl}}\right)\rightsquigarrow 2\langle\Psi_0,\mathbb{H}\rangle_{\mathcal{H}_{kl}}.$$

Equivalently,

$$\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}}-\|\Psi_0\|^2_{\mathcal{H}_{kl}}\right)\rightsquigarrow N(0,4\sigma^2),$$

where

$$\sigma^2:=\mathbb{E}_0[\langle\Psi_0,\phi_0(O)\rangle^2_{\mathcal{H}_{kl}}].$$

The scalar efficient influence function of $P\mapsto\|\Psi(P)\|^2_{\mathcal{H}_{kl}}$ at P_0 is

$$2\langle\Psi_0,\phi_0(O)\rangle_{\mathcal{H}_{kl}}.$$

Therefore $\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}}$ is regular and efficient for $\|\Psi_0\|^2_{\mathcal{H}_{kl}}$. If $\Psi_0\neq 0$, then

$$\sqrt{n}\left(\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}-\|\Psi_0\|_{\mathcal{H}_{kl}}\right)\rightsquigarrow N\left(0,\frac{\sigma^2}{\|\Psi_0\|^2_{\mathcal{H}_{kl}}}\right),$$

and the scalar efficient influence function of $P\mapsto\|\Psi(P)\|_{\mathcal{H}_{kl}}$ at P_0 is

$$\frac{\langle\Psi_0,\phi_0(O)\rangle_{\mathcal{H}_{kl}}}{\|\Psi_0\|_{\mathcal{H}_{kl}}}.$$

Proof. The first limit follows from Theorem 3. The map $T:\mathcal{H}_{kl}\to\mathbb{R}$ defined by $T(h)=\|h\|^2_{\mathcal{H}_{kl}}$ is Fréchet differentiable at every $h\in\mathcal{H}_{kl}$, with derivative

$$\dot{T}_h[u]=2\langle h,u\rangle_{\mathcal{H}_{kl}}.$$

The functional delta method therefore gives

$$\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}}-\|\Psi_0\|^2_{\mathcal{H}_{kl}}\right)\rightsquigarrow 2\langle\Psi_0,\mathbb{H}\rangle_{\mathcal{H}_{kl}}.$$

Since $\mathbb{H}$ is centered Gaussian and $\langle\Psi_0,\cdot\rangle_{\mathcal{H}_{kl}}$ is a continuous linear functional,

$$2\langle\Psi_0,\mathbb{H}\rangle_{\mathcal{H}_{kl}}\sim N(0,4\sigma^2),$$

where

$$\sigma^2=\mathbb{E}_0[\langle\Psi_0,\phi_0(O)\rangle^2_{\mathcal{H}_{kl}}].$$

The efficient influence function follows by the chain rule for pathwise differentiable Hilbert-valued parameters:

$$\phi_0^{\Theta}(O)=\dot{T}_{\Psi_0}[\phi_0(O)]=2\langle\Psi_0,\phi_0(O)\rangle_{\mathcal{H}_{kl}}.$$

Since ϕ_0 is the efficient influence function for Ψ, this scalar influence function is efficient for $P\mapsto\|\Psi(P)\|^2_{\mathcal{H}_{kl}}$.

If $\Psi_0\neq 0$, the map $S:\mathcal{H}_{kl}\to\mathbb{R}$ defined by $S(h)=\|h\|_{\mathcal{H}_{kl}}$ is Fréchet differentiable at Ψ_0, with derivative

$$\dot{S}_{\Psi_0}[u]=\frac{\langle\Psi_0,u\rangle_{\mathcal{H}_{kl}}}{\|\Psi_0\|_{\mathcal{H}_{kl}}}.$$

Another application of the functional delta method gives

$$\sqrt{n}\left(\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}-\|\Psi_0\|_{\mathcal{H}_{kl}}\right)\rightsquigarrow\frac{\langle\Psi_0,\mathbb{H}\rangle_{\mathcal{H}_{kl}}}{\|\Psi_0\|_{\mathcal{H}_{kl}}},$$

which is centered Gaussian with variance $\sigma^2/\|\Psi_0\|^2_{\mathcal{H}_{kl}}$. The efficient influence function for $P\mapsto\|\Psi(P)\|_{\mathcal{H}_{kl}}$ follows by the same chain rule. □

Delta-method implementation with reduced finite-sample bias. We contrast the plug-in V-statistic defined in Eq. (22) with its corresponding U-statistic $\widehat{Q}_U$, obtained by removing the self-pairs [Serfling, 2009].

$$\widehat{Q}_U \doteq \frac{1}{n(n-1)} \sum_{i \neq j} G_{ij}.$$

Under the assumptions of Theorem 3, we have $\widehat{Q}_U - \widehat{Q}_V = O_p(n^{-1})$. Therefore, when $\Psi_0 \neq 0$, both centers have the same delta-method limit since $\sqrt{n}(\widehat{Q}_U - \widehat{Q}_V) \rightsquigarrow 0$. In terms of asymptotic validity of the confidence sets, we can use whichever one we prefer. Since the positive bias of $\widehat{Q}_V$ is not negligible in small sample settings, we suggest centering the delta-method confidence sets on $\widehat{Q}_U$, which leads to empirically better performance. However, the same argument does not allow for a substitution of $\widehat{Q}_U$ into the triangle-inequality-based interval. At the null, its calibration uses

$$n\widehat{Q}_V \rightsquigarrow \|\mathbb{H}\|^2_{\mathcal{H}_{kl}}.$$

Hence, we can only guarantee $n\widehat{Q}_V - n\widehat{Q}_U = O_P(1)$, which is insufficient to conclude that $n\widehat{Q}_U$ has an identical distributional limit. We therefore propose to rely on $\widehat{Q}_V$ in the construction of the triangle-inequality-based interval. We believe that constructing analogous confidence sets with better small-sample behaviour and analysing them is an interesting direction for future work.

H Validity of asymptotically normal confidence sets

Notation	Meaning
$\check{\Psi}_n$	Cross-fitted one-step estimator of Ψ_0 used in the delta-method diagnostic.
$\widehat{\zeta}_n$	Bootstrap quantile for $\|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$.
$\hat{\sigma}^2$	Plug-in estimate of the delta-method variance component.
$z_{1-\beta/2}$	Standard normal quantile used for the diagnostic comparison.
$\widehat{\mathfrak{D}}_n$	Ratio of estimated quadratic remainder scale to estimated linear-term scale.

Delta method. We have the algebraic identity

$$\begin{aligned}
&\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right) \\
&= \sqrt{n}\left\langle \check{\Psi}_n - \Psi_0, \check{\Psi}_n + \Psi_0 \right\rangle_{\mathcal{H}_{kl}} \\
&= \sqrt{n}\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}} + 2\sqrt{n}\langle \check{\Psi}_n - \Psi_0, \Psi_0 \rangle_{\mathcal{H}_{kl}}.
\end{aligned}$$

We refer to $2\sqrt{n}\langle \check{\Psi}_n - \Psi_0, \Psi_0 \rangle_{\mathcal{H}_{kl}}$ as the *linear term* and $\sqrt{n}\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}}$ as the *quadratic remainder*. For a fixed $\Psi_0 \neq 0$, we derive a distributional limit for $\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right)$. Suppose the conclusion of Theorem 3 holds, namely

$$\sqrt{n}(\check{\Psi}_n - \Psi_0) \rightsquigarrow \mathbb{H}.$$

By the Continuous Mapping Theorem, $n\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}} \rightsquigarrow \|\mathbb{H}\|^2_{\mathcal{H}_{kl}}$. Thus $\sqrt{n}\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}} \rightsquigarrow 0$. By another application of the Continuous Mapping Theorem, $\sqrt{n}\left\langle \check{\Psi}_n - \Psi_0, \Psi_0 \right\rangle \rightsquigarrow \langle \mathbb{H}, \Psi_0 \rangle_{\mathcal{H}_{kl}}$. Thus

$$\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right) \rightsquigarrow 2\langle \mathbb{H}, \Psi_0 \rangle_{\mathcal{H}_{kl}}.$$

Since $2\langle \mathbb{H}, \Psi_0 \rangle_{\mathcal{H}_{kl}} \sim \mathcal{N}(0, 4\mathbb{E}[\langle \phi_0(O), \Psi_0 \rangle^2_{\mathcal{H}_{kl}}])$, this yields a closed form asymptotic variance that can be used to construct a calibrated confidence interval for $\sqrt{n}\left(\|\check{\Psi}_n\|^2_{\mathcal{H}_{kl}} - \|\Psi_0\|^2_{\mathcal{H}_{kl}}\right)$.

However, in finite samples, the quality of the approximation heavily depends on the size of $\sqrt{n}\|\check{\Psi}_n - \Psi_0\|^2_{\mathcal{H}_{kl}}$ relative to $2\sqrt{n}\langle \check{\Psi}_n - \Psi_0, \Psi_0 \rangle_{\mathcal{H}_{kl}}$. We therefore propose a data-driven diagnostic via the

bootstrapped empirical measures $\mathbb{P}_n^{(b)}$. Recall in Section 3, we defined $\widehat{\zeta}_n$ as the $(1-\alpha)$-quantile of $\|\mathbb{H}_n^{(b)}\|_{\mathcal{H}_{kl}}^2$, conditionally on the original sample $(O_1, \ldots, O_n)$. The confidence set $\mathcal{C}_n(\widehat{\zeta}_n)$ contains 0 if and only if $n\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^2 \leq \widehat{\zeta}_n$.

If the linear term has reached its asymptotic regime, then

$$2\sqrt{n}\left\langle \check{\Psi}_n - \Psi_0, \Psi_0 \right\rangle_{\mathcal{H}_{kl}} \overset{d}{\approx} N(0, 4\sigma^2), \qquad \sigma^2 = \mathrm{Var}\left(\langle \Psi_0, \mathbb{H} \rangle_{\mathcal{H}_{kl}}\right).$$

Thus the $(1-\beta)$ quantile of the absolute value of the linear term is approximated by

$$2\hat{\sigma} z_{1-\beta/2},$$

where $\hat{\sigma}^2$ is the plug-in variance estimator from Proposition 3. On the other hand, since $\hat{\zeta}_n$ estimates the $(1-\alpha)$ quantile of $\|\mathbb{H}\|_{\mathcal{H}_{kl}}^2$, the quantity $n^{-1/2}\hat{\zeta}_n$ estimates the $(1-\alpha)$ quantile of the quadratic remainder

$$\sqrt{n}\|\check{\Psi}_n - \Psi_0\|_{\mathcal{H}_{kl}}^2.$$

This gives the diagnostic

$$\widehat{\mathfrak{D}}_n = \frac{n^{-1/2}\hat{\zeta}_n}{2\hat{\sigma} z_{1-\beta/2}}.$$

A simple warning sign is that the triangle-inequality-based confidence set (Eq. (5)) contains zero:

$$0 \in \mathcal{C}_n(\hat{\zeta}_n) \quad \Longleftrightarrow \quad n\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}^2 \leq \hat{\zeta}_n.$$

I Cross U-statistics and confidence intervals

This section will be deliberately informal. We will explore a sample splitting construction closely related to cross U-statistics [Kim and Ramdas, 2024, Shekhar et al., 2023]. A cross U-statistic uses sample splitting to obtain a test-statistic that is valid under both the null and away from the null, offering the possibility for a CLT-based debiased test, obviating the need for computationally intensive bootstrap sampling. Recall in the main text, we proposed two confidence interval (CI) constructions. The first is based on the reverse triangular inequality (RTI), which is exact under the null, pointwise valid everywhere, and locally uniformly valid under local alternatives $\Psi_n = \frac{h}{\sqrt{n}}$. The second is based on the delta method (DM), which is degenerate under the null but efficient away from the null. We recall (informally) the asymptotic linearity expansion. Suppose along a quadratic mean differentiable (QMD) sequence P_n, we have

$$\widehat{\Psi}_n = \Psi_n + \frac{\mathbb{H}_n}{\sqrt{n}} + o_p(1/\sqrt{n}),$$

where $\mathbb{H}_n \rightsquigarrow \mathbb{H}$, and $\mathbb{H}$ is a Gaussian element in the RKHS. The RTI CI's validity under local alternatives is due to the regularity of $\widehat{\Psi}_n$. However, from the identity

$$\|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}}^2 - \|\Psi\|_{\mathcal{H}_{kl}}^2 = 2\langle \Psi, \widehat{\Psi}_n - \Psi \rangle_{\mathcal{H}_{kl}} + \|\widehat{\Psi}_n - \Psi\|_{\mathcal{H}_{kl}}^2,$$

we see that using $\widehat{\Psi}_n - \Psi_n = O_p(1/\sqrt{n})$, along local alternatives Ψ_n, the two terms on the right are of the same order, and we have

$$n\left(\|\widehat{\Psi}_n\|_{\mathcal{H}_{kl}}^2 - \|\Psi_n\|_{\mathcal{H}_{kl}}^2\right) \rightsquigarrow 2\langle h, \mathbb{H} \rangle_{\mathcal{H}_{kl}} + \|\mathbb{H}\|_{\mathcal{H}_{kl}}^2.$$

Hence the CLT-based DM CI need not cover along local alternatives, in other words, it does not achieve locally uniform coverage. The importance of locally uniform coverage can be understood as follows. In finite samples, we may not have enough data to tell whether we are at P_0 or at a neighbouring law $\widetilde{P}_0$, which is close to P_0 in Hellinger distance. A procedure that does not achieve locally uniform coverage could badly undercover in finite samples compared with one which does achieve it.

We now show that by using a sample-splitting construction, we can obtain a CLT-based CI which is valid everywhere but ***conservative*** at the null. Furthermore, we show that this procedure also has a

non-Gaussian limit along local alternatives, therefore it also fails to attain locally uniform coverage. We split the testing sample into two halves I_1 and I_2, with $|I_1| = |I_2| = m$. We define

$$\widehat{\Psi}_A = \frac{1}{m}\sum_{i\in I_1}\widehat{D}_i, \quad \widehat{\Psi}_B = \frac{1}{m}\sum_{j\in I_2}\widehat{D}_j.$$

The cross U-statistic is $\widehat{\theta} \doteq \langle \widehat{\Psi}_A, \widehat{\Psi}_B\rangle_{\mathcal{H}_{kl}}$. We suppose the following expansions hold:

$$\widehat{\Psi}_A = \Psi_m + \frac{G_A}{\sqrt{m}} + o_p(1/\sqrt{m})$$
$$\widehat{\Psi}_B = \Psi_m + \frac{G_B}{\sqrt{m}} + o_p(1/\sqrt{m}),$$

where G_A and G_B are independent mean-zero Gaussian elements in $\mathcal{H}_{kl}$ with covariance operator Σ. We write $\delta_A \doteq \widehat{\Psi}_A - \Psi$ and $\delta_B \doteq \widehat{\Psi}_B - \Psi$. Then the following decomposition holds:

$$\widehat{\theta} - \theta = \langle \Psi + \delta_A, \Psi + \delta_B\rangle_{\mathcal{H}_{kl}} - \|\Psi\|^2_{\mathcal{H}_{kl}} = \langle \delta_A + \delta_B, \Psi\rangle_{\mathcal{H}_{kl}} + \langle \delta_A, \delta_B\rangle_{\mathcal{H}_{kl}}.$$

We now outline why a studentized test statistic derived from $\widehat{\theta} - \theta$ obeys a CLT. We will also show the studentizer is constructed differently under the null and the alternative. We assume nuisances are fit on an independent split, and we condition on that split throughout.

First consider the null $H_0 : \Psi = 0$. Define the σ-algebra

$$\mathcal{F}_B = \sigma\left(\widehat{D}_j : j \in I_2\right).$$

We compute the conditional variance

$$\begin{aligned}\mathrm{Var}\left(\sqrt{m}\widehat{\theta} \mid \mathcal{F}_B\right) &= \mathrm{Var}\left(\frac{1}{\sqrt{m}}\sum_{i\in I_1}\langle \widehat{D}_i, \widehat{\Psi}_B\rangle_{\mathcal{H}_{kl}} \mid \mathcal{F}_B\right) \\ &= \mathrm{Var}\left(\langle \widehat{D}_i, \widehat{\Psi}_B\rangle_{\mathcal{H}_{kl}} \mid \mathcal{F}_B\right).\end{aligned}$$

An empirical variance estimator is

$$\widehat{s}^2_B = \frac{1}{m}\sum_{i\in I_1}\left\{\langle \widehat{D}_i, \widehat{\Psi}_B\rangle_{\mathcal{H}_{kl}} - \widehat{\theta}\right\}^2.$$

Using a Lyapunov CLT, one has

$$\frac{\sqrt{m}(\widehat{\theta} - \theta)}{\sqrt{\frac{1}{m}\sum_{i\in I_1}\{\langle \widehat{D}_i, \widehat{\Psi}_B\rangle_{\mathcal{H}_{kl}} - \widehat{\theta}\}^2}} \rightsquigarrow \mathcal{N}(0,1),$$

whereas under the alternative $\Psi \neq 0$, one has

$$\sqrt{m}(\widehat{\theta} - \theta) = \langle \Psi, \sqrt{m}\delta_A\rangle_{\mathcal{H}_{kl}} + \langle \Psi, \sqrt{m}\delta_B\rangle_{\mathcal{H}_{kl}} + \langle \sqrt{m}\delta_A, \delta_B\rangle_{\mathcal{H}_{kl}}.$$

The first two terms dominate the expression, as the third one is of order $o_p(1)$. We have

$$\mathrm{Var}(\langle \Psi, \sqrt{m}\delta_A\rangle_{\mathcal{H}_{kl}}) = \mathrm{Var}(\langle \Psi, \sqrt{m}\delta_B\rangle_{\mathcal{H}_{kl}}).$$

By independence, we have

$$\mathrm{Var}(\sqrt{m}(\widehat{\theta} - \theta)) = 2\mathrm{Var}(\langle \Psi, \sqrt{m}\delta_A\rangle_{\mathcal{H}_{kl}}) + o_p(1).$$

Additionally, we have

$$\begin{aligned}&\mathrm{Var}\left(\langle \widehat{\Psi}_B, \sqrt{m}\delta_A\rangle_{\mathcal{H}_{kl}} \mid \mathcal{F}_B\right) \\ &= \mathrm{Var}\left(\left\langle \widehat{\Psi}_B, \frac{1}{\sqrt{m}}\sum_{i\in I_1}\widehat{D}_i\right\rangle_{\mathcal{H}_{kl}} \mid \mathcal{F}_B\right) \\ &= \mathrm{Var}\left(\langle \widehat{\Psi}_B, \widehat{D}_i\rangle_{\mathcal{H}_{kl}} \mid \mathcal{F}_B\right).\end{aligned}$$

Therefore, a consistent empirical variance estimator is $2\widehat{s}_B^2$. Using a Lyapunov CLT, we can show that

$$\frac{\sqrt{m}(\widehat{\theta} - \theta)}{\sqrt{\frac{2}{m}\sum_{i \in I_1}\{\langle \widehat{D}_i, \widehat{\Psi}_B \rangle_{\mathcal{H}_{kl}} - \widehat{\theta}\}^2}} \rightsquigarrow \mathcal{N}(0, 1).$$

We therefore inflate the variance by a factor of 2, or the standard error by a factor of $\sqrt{2}$, for the alternative. Naturally, we can use the inflated variance CI both under the null and the alternative, which would be conservative under the null. We now show that the studentized cross U-statistic is not asymptotically normal under local alternatives, using the natural studentizer $\widehat{s}_B$. Under $\Psi_m = \frac{h}{\sqrt{m}}$, with $\theta_m = \|\Psi_m\|^2_{\mathcal{H}_{kl}}$, we have

$$\begin{aligned} m\widehat{s}_B^2 &\rightsquigarrow \langle h + G_B, \Sigma(h + G_B) \rangle_{\mathcal{H}_{kl}}, \\ m(\widehat{\theta} - \theta_m) &= \langle G_A + G_B, \sqrt{m}\Psi_m \rangle_{\mathcal{H}_{kl}} + \langle G_A, G_B \rangle_{\mathcal{H}_{kl}} + o_p(1) \\ &\rightsquigarrow \langle h, G_A + G_B \rangle_{\mathcal{H}_{kl}} + \langle G_A, G_B \rangle_{\mathcal{H}_{kl}}. \end{aligned}$$

Therefore the centered studentized statistic $\sqrt{m}(\widehat{\theta} - \theta_m)/\widehat{s}_B$ has limit

$$\frac{\langle h, G_A + G_B \rangle_{\mathcal{H}_{kl}} + \langle G_A, G_B \rangle_{\mathcal{H}_{kl}}}{\sqrt{\langle h + G_B, \Sigma(h + G_B) \rangle_{\mathcal{H}_{kl}}}}$$

under the local alternatives. By comparison, Shekhar et al. [2023] use local alternatives to study local power but do not establish local uniform coverage of a CLT-based CI in their setting.

I.1 Empirical evaluation

We compare the Wald interval based on the cross U-statistic above with the debiased delta-method interval and the interval based on the reverse triangle inequality in Appendix G. We use the synthetic X-dependent residual experiment in Appendix J.1. All three procedures target $\|\Psi\|_{\mathcal{H}_{kl}} = \sqrt{\mathrm{HSIC}(X, \xi_0(W, Y))}$ and use the same fitted nuisance functions.

In these experiments, we reuse the cross-fitted one-step summands from Appendix F. For observation O_i, let $k(i)$ denote its evaluation fold and let $\widehat{D}_i = \widehat{D}^{-k(i)}(O_i) \in \mathcal{H}_{kl}$, with the nuisance functions and centering terms fitted outside that fold. Independently of the observations, we partition each evaluation fold into two equal parts and take their respective unions to form I_1 and I_2. Thus each half contains $m = n/2$ observations and half of every evaluation fold. Using the notation above, we compute

$$\widehat{\Psi}_A = \frac{1}{m}\sum_{i \in I_1} \widehat{D}_i, \qquad \widehat{\Psi}_B = \frac{1}{m}\sum_{j \in I_2} \widehat{D}_j.$$

The Gram entries $G_{ij} = \langle \widehat{D}_i, \widehat{D}_j \rangle_{\mathcal{H}_{kl}}$ give

$$\begin{aligned} \widehat{\theta} &= \langle \widehat{\Psi}_A, \widehat{\Psi}_B \rangle_{\mathcal{H}_{kl}} = \frac{1}{m^2}\sum_{i \in I_1}\sum_{j \in I_2} G_{ij}, \\ \widehat{s}_B^2 &= \frac{1}{m}\sum_{i \in I_1}\left\{\frac{1}{m}\sum_{j \in I_2} G_{ij} - \widehat{\theta}\right\}^2. \end{aligned}$$

We apply the finite-sample variance correction $m/(m-1)$ to $\widehat{s}_B^2$. We obtain the variance-inflated confidence set

$$\mathcal{I}_\times = \left\{\sqrt{t} : t \geq 0, \quad |\widehat{\theta} - t| \leq z_{0.975}\sqrt{\frac{2}{m-1}}\,\widehat{s}_B\right\}, \tag{24}$$

where $z_{0.975}$ is the 0.975 quantile of the standard normal distribution. We construct the debiased delta-method interval by centering the squared-norm Wald interval at the U-statistic in Appendix G,

intersecting it with $[0, \infty)$, and taking square roots. For the interval based on the reverse triangle inequality, we use the norm of the full cross-fitted estimator,

$$\|\check{\Psi}_n\|_{\mathcal{H}_{kl}}, \qquad \check{\Psi}_n = \frac{1}{2}(\widehat{\Psi}_A + \widehat{\Psi}_B),$$

and the bootstrap quantile from Appendix G.

We fix $n = 250$, $K = 5$, and mean smoothness $s_m = 1$, and vary noise smoothness $s_\epsilon \in \{0.25, 0.5, 0.75\}$ and heterogeneity strength $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Each nuisance fit uses 200 training observations, and each of I_1 and I_2 contains 125 observations. Each setting uses 100 paired simulation repetitions, with the nuisance-fitting choices and kernel bandwidths from Appendix J.1. Within each repetition, all three methods use the same fitted nuisances.

Table 1 reports mean width on the $\|\Psi\|_{\mathcal{H}_{kl}}$ scale and the empirical probability of including zero, at nominal confidence level 95%. At $\rho = 0$, the probability of including zero is the coverage probability; at $\rho > 0$, its complement is the probability of rejecting residual independence. Across all reported settings, the Cross U-statistic interval has a larger mean width than the delta-method interval and a smaller mean width than the interval based on the reverse triangle inequality. It includes zero more frequently than the latter interval.

Table 1: Mean width and empirical probability of including zero (in parentheses) for nominal 95% confidence sets for $\|\Psi\|_{\mathcal{H}_{kl}}$, with $n = 250$ and $s_m = 1$. All methods are debiased and use five-fold cross-fitting and the same 100 repetitions per setting.

Method	$\rho = 0.0$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$
			(a) $s_\epsilon = 0.25$			
Cross U-statistic	0.0411 (0.97)	0.0421 (0.96)	0.0450 (0.97)	0.0491 (0.96)	0.0538 (0.95)	0.0580 (0.92)
delta method	0.0339 (0.96)	0.0351 (0.96)	0.0388 (0.97)	0.0430 (0.96)	0.0480 (0.96)	0.0521 (0.95)
RTI	0.0765 (0.93)	0.0780 (0.92)	0.0816 (0.89)	0.0868 (0.85)	0.0928 (0.81)	0.0996 (0.72)
			(b) $s_\epsilon = 0.5$			
Cross U-statistic	0.0411 (0.97)	0.0428 (0.97)	0.0472 (0.97)	0.0536 (0.95)	0.0590 (0.89)	0.0616 (0.75)
delta method	0.0339 (0.96)	0.0361 (0.95)	0.0412 (0.95)	0.0482 (0.95)	0.0540 (0.89)	0.0587 (0.83)
RTI	0.0765 (0.93)	0.0785 (0.90)	0.0831 (0.83)	0.0893 (0.68)	0.0951 (0.63)	0.1015 (0.53)
			(c) $s_\epsilon = 0.75$			
Cross U-statistic	0.0411 (0.97)	0.0435 (0.97)	0.0498 (0.96)	0.0570 (0.87)	0.0618 (0.77)	0.0647 (0.65)
delta method	0.0339 (0.96)	0.0372 (0.97)	0.0445 (0.94)	0.0532 (0.90)	0.0589 (0.80)	0.0619 (0.65)
RTI	0.0765 (0.93)	0.0791 (0.89)	0.0847 (0.75)	0.0910 (0.58)	0.0962 (0.44)	0.1014 (0.34)

J Experiment details

Code and instructions for reproducing the experiments are available at https://github.com/Wornbard/EfficientAdditiveNoiseHSIC.

Notation	Meaning
n	Sample size used in experiment figures and tables.
s_m, s_ϵ	Parameters controlling the decay of the Fourier coefficients of the mean and log noise scale.
ρ	Strength of residual heterogeneity; $\rho = 0$ is homogeneous noise.
$a_k, b_k, \phi_{m,k}, \phi_{\sigma,k}$	Random Fourier coefficients and phases in the synthetic data-generating mechanisms.
$c_m, c_\epsilon, \sigma_0$	Scaling constants for the mean/noise series and the baseline noise scale.
$g(x), h(x), c$	Auxiliary functions and scale constant used in heteroscedastic noise examples.
$m_{Y\|X}, m_{X\|Y}$	Direction-specific regressions in the causal-arrow experiment.
$T, \sigma(0), \sigma(1)$	Extra covariate and group-specific residual scales in the covariate-adjusted experiment.

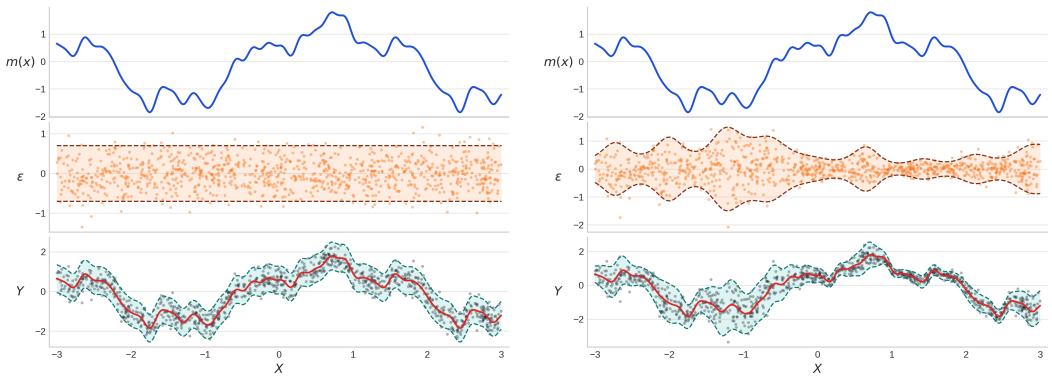
We use five-fold cross-fitting in the synthetic calibration, causal-direction, MSE and cross-U experiments, and four-fold cross-fitting in the covariate-adjusted experiment. The nuisance functions $\hat{m}$ and $\hat{v}$ are fitted by kernel ridge regression (vector-valued where appropriate). We found that Matérn kernels perform slightly better than smoother alternatives at learning v. Permutation tests and bootstrap confidence sets use 1,000 resamples in the synthetic calibration and MSE experiments, and 399 in the causal-direction experiments, including the classical comparison. MSE-summary uncertainty intervals use 10,000 paired resamples of simulation repetitions. All experiments can be performed on a single machine with a single NVIDIA RTX 4500 GPU and 64GB of RAM. To accelerate the evaluation of our method across settings and repetitions, we parallelised the computations across multiple machines. Each one had 64GB of RAM, and multiple different GPUs were used, but the GPU capacity is not a core bottleneck. We did not evaluate how quickly our procedures can be run on a CPU. On the aforementioned RTX 4500 GPU, fitting a 5-fold cross-fitted confidence set in $d_x = d_y = 1$ and with 1000 data points took on the order of 1 to 2 minutes. For fixed outcome dimension and a fixed number of folds, storing the dense kernel and derivative matrices requires $O(n^2)$ memory. Forming their entries uses $O(n^2)$ kernel evaluations; a dense kernel ridge regression solve takes $O(n^3)$ arithmetic operations. We did not investigate ways of mitigating this empirically; however, since the implementation is expressed through Gram matrices and their derivative analogues, standard kernel-acceleration ideas such as random Fourier features, Nyström approximations, and incomplete or block U-statistics are natural candidates for future scalability work [Rahimi and Recht, 2007, Williams and Seeger, 2000, Zhang et al., 2018, Schrab et al., 2022].

J.1 Datasets used in the synthetic calibration-evaluation experiments

This subsection gives the scalar data-generating mechanism used for the reported calibration-evaluation experiments. The same random Fourier construction can be extended componentwise to multivariate X or Y, but the reported main experiments keep $d_x = d_y = 1$ to focus on residual-estimation difficulty.

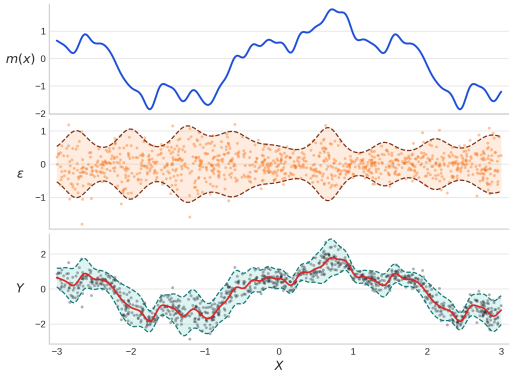
The data-generating process is

$$Y = m(X) + \sigma(X)Z,$$



(a) $s_m = 1$, $\rho = 0$

(b) $s_m = 1$, $\rho = 0.5$, $s_\epsilon = 0.75$



(c) $s_m = 1$, $\rho = 0.3$, $s_\epsilon = 0.25$

Figure 5: Example synthetic datasets. Panel (a) is a homogeneous-noise null setting; panels (b) and (c) are alternatives with X-dependent noise scale.

where $Z \sim \mathcal{N}(0,1)$ is sampled independently of X. The structural function used to construct the synthetic datasets is

$$m(x) = c_m \sum_{k=1}^{20} a_k k^{-s_m} \sin(kx + \phi_{m,k}),$$

with $a_k \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\phi_{m,k} \overset{i.i.d.}{\sim} \text{Uniform}[-\pi, \pi]$, and c_m chosen post-hoc so that $\text{Var}(m(X)) = 1$. Similarly,

$$g(x) = \sum_{k=1}^{20} b_k k^{-s_\epsilon} \sin(kx + \phi_{\sigma,k}),$$
$$\sigma(x) = \sigma_0 \exp(\rho c_\epsilon g(x)),$$

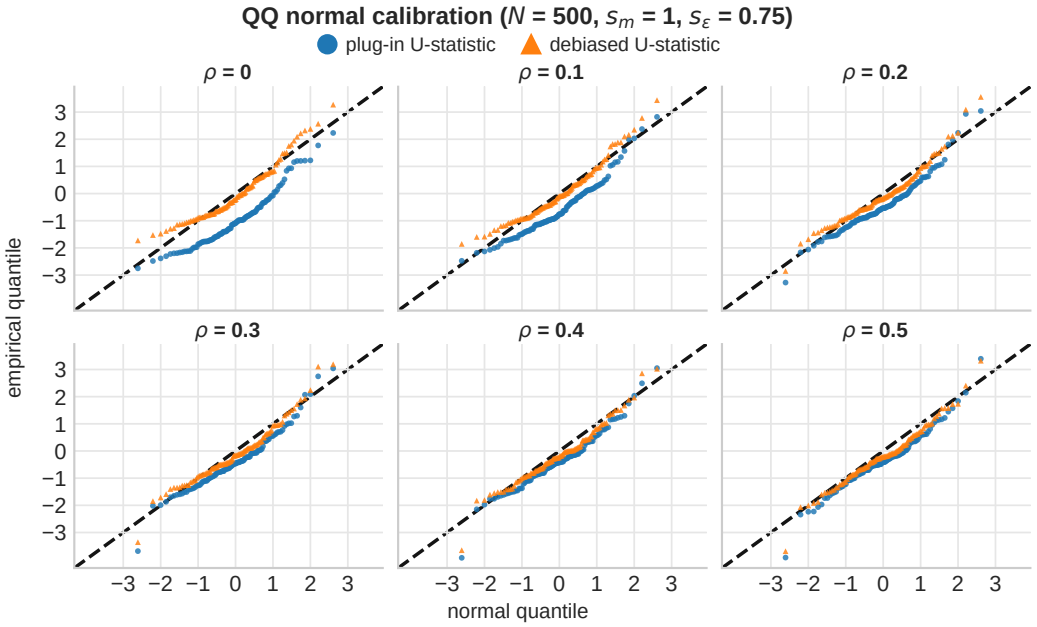
where $\rho \geq 0$ (a value of 0 corresponds to homogeneous noise with standard deviation σ_0). Weights b_k and phases $\phi_{\sigma,k}$ are sampled as above. c_ϵ is chosen to ensure $\text{Var}(g(X)c_\epsilon) = 1$, and in all our experiments $\sigma_0 = 0.35$.

The HSIC kernels are Gaussian with bandwidths tuned by the median heuristic, applied to full X for the $\mathcal{H}_k$ kernel, and $\hat{\xi}$, the estimated residuals, on the $\mathcal{H}_l$ kernel. Example datasets are shown in Figure 5. We use the classical permutation HSIC test [Gretton et al., 2007] as the plug-in baseline because it is the canonical residual-independence diagnostic in ANM experiments. Improvements to the underlying raw independence test, such as Schrab et al. [2022], are unrelated to the debiasing problem studied here, though it would be an interesting direction for future work to extend our estimation methodology to functionals that aggregate over cross-covariance operators in multiple RKHSs.

J.2 Normal approximations and confidence interval coverage

We report empirical coverage for the nominal 95% confidence intervals in the same synthetic setting used for the main calibration plot, namely $s_m = 1$ and $s_\epsilon = 0.75$. The intervals based on the reverse triangle inequality (RTI) are valid at the null and remain conservative under alternatives. This is visible in Table 2: coverage is close to nominal when $\rho = 0$, and approaches 100% as ρ increases. The delta-method intervals are not guaranteed to be valid at $\rho = 0$, where the delta method does not apply, but under alternatives they are much less conservative. For the debiased estimator, Table 3 shows finite-sample undercoverage in several settings. Throughout the reported delta comparisons, we use the U-statistic as the center of the squared-norm Wald interval, intersect that interval with $[0, \infty)$, and take square roots.

The QQ plots in Figure 6 tell the same story. Away from $\rho = 0$, the debiased statistic tracks the Gaussian reference line reasonably well, especially in the central quantiles. The plug-in statistic exhibits visible bias and slope distortions, while the debiased statistic is substantially better aligned with the target. At $\rho = 0$, the norm map is not differentiable at the target operator, and the squared-norm map has zero first derivative. The Gaussian approximation derived for a fixed alternative therefore gives way to the quadratic limit described in Appendix I.



Note. At $\rho = 0$ there is no asymptotic normality under the null, so the QQ plot is not expected to fit the Gaussian reference line in that panel.

Figure 6: Quantile–quantile plots of the plug-in and debiased HSIC U-statistics for the X-dependent residual experiment at $n = 500$, across $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

Table 2: Coverage of the RTI confidence intervals for $s_m = 1$ and $s_\epsilon = 0.75$. Parentheses give 95% error bars.

ρ	$n = 250$	$n = 500$	$n = 1000$
0	0.930 (-0.050, +0.050)	0.955 (-0.039, +0.039)	0.916 (-0.053, +0.053)
0.1	0.970 (-0.033, +0.030)	0.982 (-0.025, +0.018)	0.991 (-0.018, +0.009)
0.2	0.990 (-0.020, +0.010)	0.991 (-0.018, +0.009)	1.000 (-0.000, +0.000)
0.3	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)
0.4	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)
0.5	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)	1.000 (-0.000, +0.000)

Table 3: Coverage of the debiased delta-method confidence intervals for $s_m = 1$ and $s_\epsilon = 0.75$. The delta-method interval is only reported for $\rho > 0$. Parentheses give 95% error bars.

ρ	$n = 250$	$n = 500$	$n = 1000$
0.1	0.940 (-0.047, +0.047)	0.936 (-0.046, +0.046)	0.898 (-0.057, +0.057)
0.2	0.900 (-0.059, +0.059)	0.927 (-0.049, +0.049)	0.909 (-0.054, +0.054)
0.3	0.910 (-0.056, +0.056)	0.945 (-0.042, +0.042)	0.935 (-0.046, +0.046)
0.4	0.860 (-0.068, +0.068)	0.936 (-0.046, +0.046)	0.962 (-0.036, +0.036)
0.5	0.900 (-0.059, +0.059)	0.936 (-0.046, +0.046)	0.936 (-0.046, +0.046)

J.3 Causal arrow experiment details

The evaluated dataset is a simple example similar to the ones evaluated in the synthetic experiments of X-dependent residuals in Section 4. The structural function is

$$m(x) = 0.55\,x + 0.95\,\sin(1.65\,x) + 0.38\,\sin(3.6\,x) + 0.1\,x^3.$$

The conditional standard deviation of the normally distributed noise is

$$\begin{aligned} \sigma(x) &= 0.45\exp(c\,\rho\{h(x) - \mathbb{E}[h(X)]\}), \\ h(x) &= \sin(1.6\,x) + 0.35\cos(2.4\,x), \\ c &= \sqrt{1/\mathrm{Var}(h(X))}. \end{aligned}$$

We observe that the only procedures that are properly calibrated under the null are the debiased bootstrap procedure and the full-data cross-fitted permutation test, which was in fact very conservative, with its p-value distribution far from the uniform law that would correspond to perfect calibration. Among the tests that correctly fail to reject the null, only ours and the full-sample cross-fitted permutation baseline attain non-trivial power, with the former performing significantly better. The split-data tests that were calibrated attain very poor power under the alternative because the testing sample is too small.

J.4 Covariate-adjusted residual diagnostics

Each simulated dataset has $n = 120$ observations, with $X \sim \mathrm{Bernoulli}(1/2)$, $T \sim \mathrm{Unif}[-\pi, \pi]$, and

$$Y = m(X, T) + \sigma(X)\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1).$$

The mean is

$$m(x, t) = 1.5\left\{0.2x + \sum_{k=1}^{5}\left[\frac{0.30}{k}\sin(kt) + \frac{0.25}{k}\cos(kt) + \frac{0.22}{k}x\sin(kt) + \frac{0.18}{k}x\cos(kt)\right]\right\}.$$

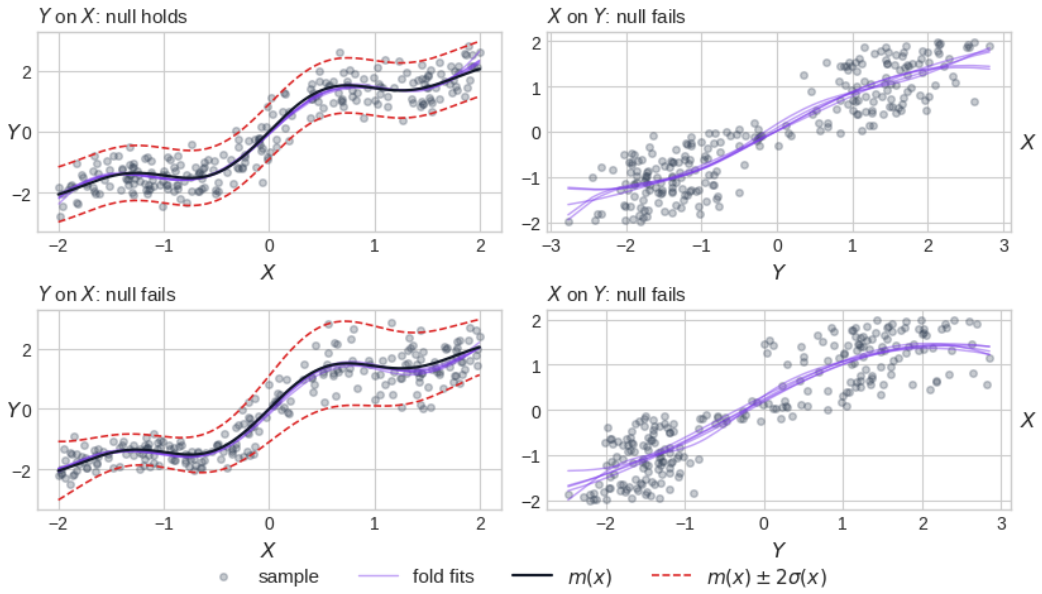


Figure 7: Example data from the causal ANM experiment, where the additive-noise direction is $X \to Y$.

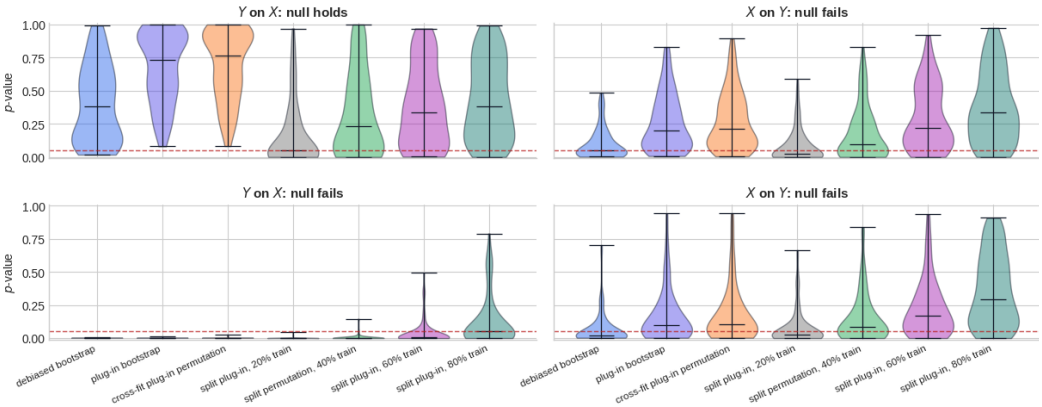


Figure 8: Distribution of causal-arrow scores across repetitions of the two-way goodness-of-fit experiment.

The null uses $\sigma(0) = \sigma(1) = 0.45$, while the alternative uses $\sigma(0) = 0.35, \sigma(1) = 0.55$.

The experiment uses four-fold cross-fitting and kernel ridge regression in the fixed 22-dimensional trigonometric feature basis. Ridge penalties are selected by three-fold cross-validation within each training fold. All methods use the same 16 fixed random Fourier features for the residual kernel, with Gaussian lengthscale 1. Figure 3 uses 500 alternative repetitions, retaining empty confidence sets in coverage and mean width.

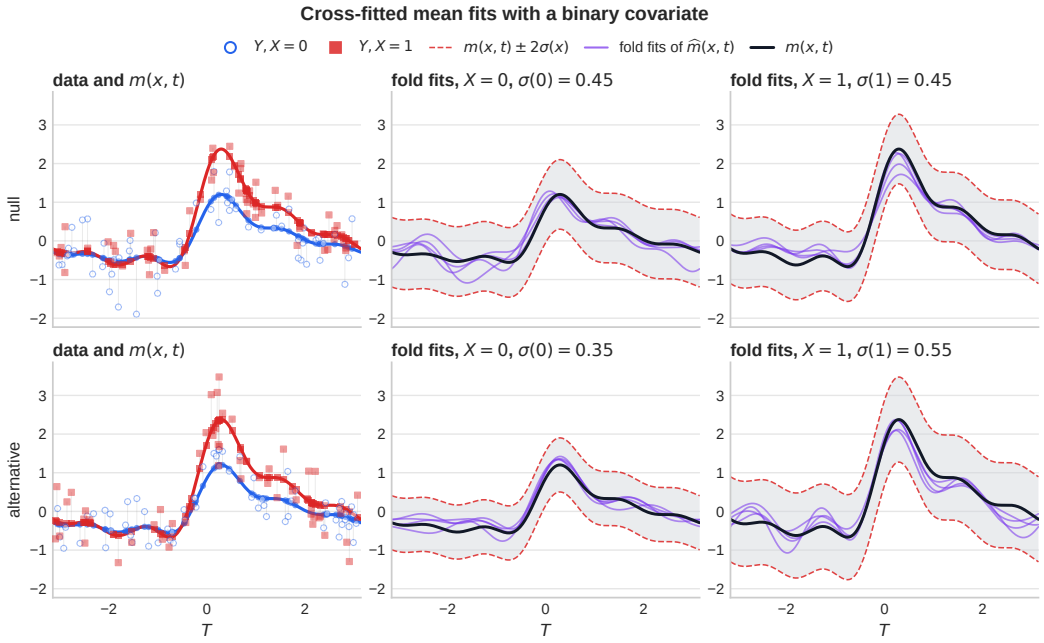


Figure 9: Fold-specific structural function fits in the covariate-adjusted residual experiment.

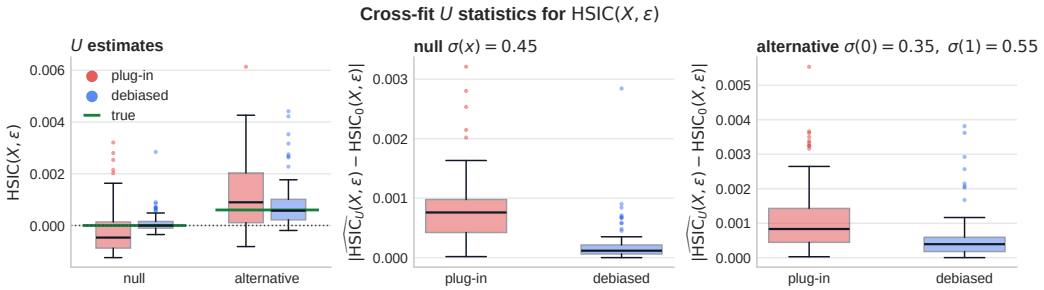


Figure 10: Comparison of the feasible U-statistic estimator with the oracle target in the covariate-adjusted residual experiment.

J.5 Ranges of empirical coverage point estimates

We report ranges of observed empirical coverage means for intervals based on the reverse triangle inequality (RTI) and the debiased delta method across the indicated noise-dependence levels.

Table 4: Ranges of empirical coverage point estimates over ρ for the remaining synthetic calibration settings at $n = 250$, excluding the main setting $s_m = 1$ and $s_\epsilon = 0.75$. The RTI range is over $\rho \in \{0, 0.1, \ldots, 0.5\}$; the delta-method range is over $\rho \in \{0.1, \ldots, 0.5\}$.

s_m	s_ϵ	RTI	Debiased delta method
1.0	0.25	0.940–0.990	0.910–0.940
1.0	0.5	0.930–1.000	0.870–0.950
1.5	0.25	0.920–1.000	0.900–0.930
1.5	0.5	0.890–1.000	0.890–0.920
1.5	0.75	0.890–1.000	0.890–0.910
2.0	0.25	0.920–1.000	0.890–0.950
2.0	0.5	0.910–0.990	0.870–0.930
2.0	0.75	0.910–1.000	0.880–0.910

Table 5: Ranges of empirical coverage point estimates over ρ for the remaining synthetic calibration settings at $n = 500$. The entries are ranges of observed means, not confidence intervals.

s_m	s_ϵ	RTI	Debiased delta method
1.0	0.25	0.938–1.000	0.909–0.936
1.0	0.5	0.955–1.000	0.918–0.936
1.5	0.25	0.933–1.000	0.927–0.955
1.5	0.5	0.964–1.000	0.918–0.945
1.5	0.75	0.964–1.000	0.918–0.945
2.0	0.25	0.929–1.000	0.927–0.955
2.0	0.5	0.982–1.000	0.927–0.945
2.0	0.75	0.982–1.000	0.936–0.945

Table 6: Ranges of empirical coverage point estimates over ρ for the remaining synthetic calibration settings at $n = 1000$. The entries are ranges of observed means, not confidence intervals.

s_m	s_ϵ	RTI	Debiased delta method
1.0	0.25	0.918–0.991	0.897–0.927
1.0	0.5	0.918–1.000	0.889–0.945
1.5	0.25	0.914–0.991	0.907–0.936
1.5	0.5	0.896–1.000	0.909–0.936
1.5	0.75	0.907–1.000	0.923–0.945
2.0	0.25	0.895–1.000	0.914–0.945
2.0	0.5	0.879–1.000	0.913–0.954
2.0	0.75	0.882–1.000	0.925–0.954

J.6 Additional QQ calibration plots at $n = 500$

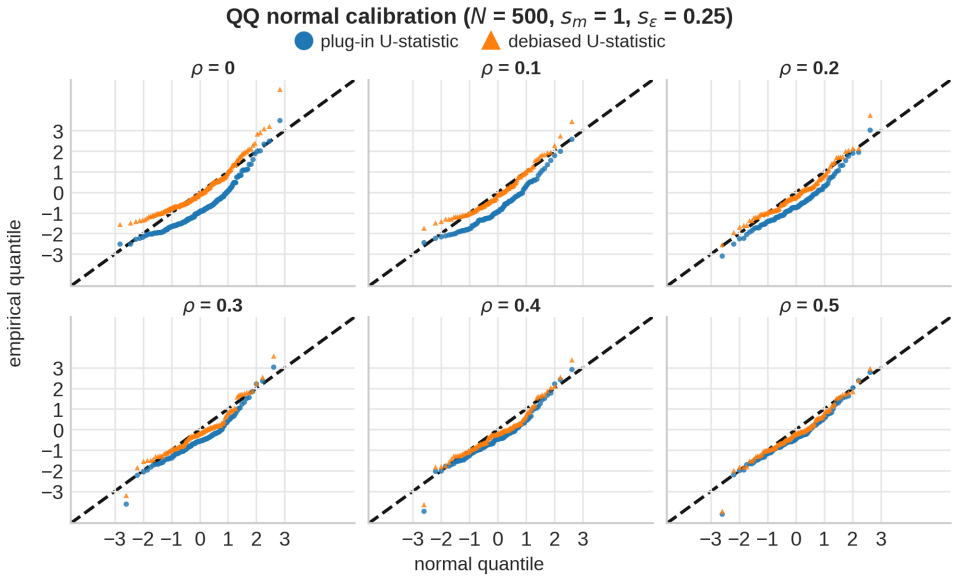


Figure 11: QQ calibration at $n = 500$, $s_m = 1$, and $s_\epsilon = 0.25$.

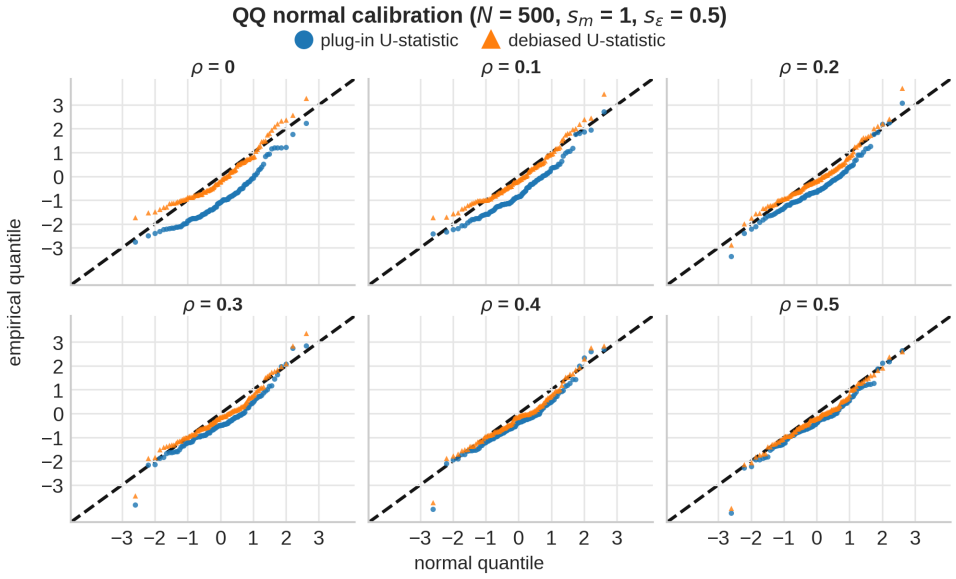


Figure 12: QQ calibration at $n = 500$, $s_m = 1$, and $s_\epsilon = 0.5$.

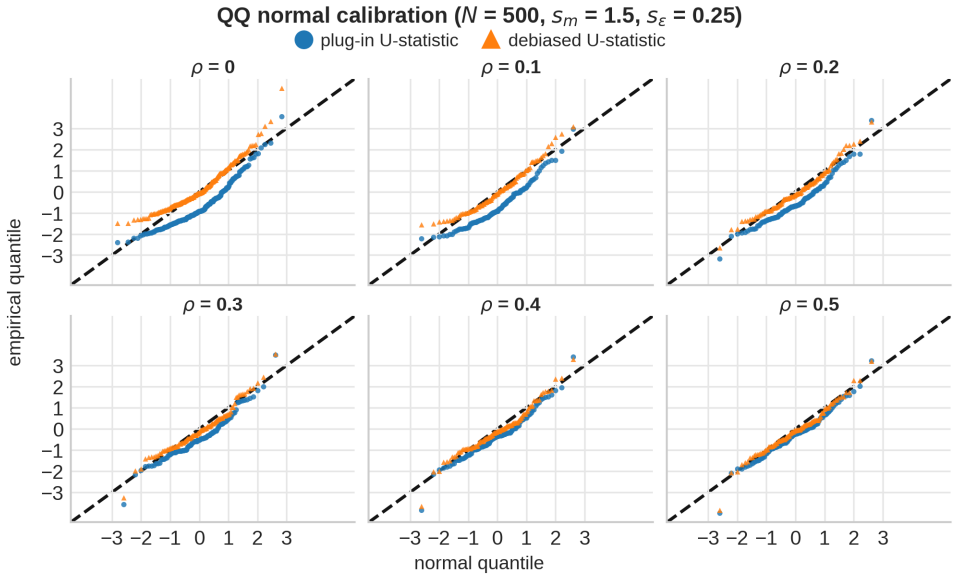


Figure 13: QQ calibration at $n = 500$, $s_m = 1.5$, and $s_\epsilon = 0.25$.

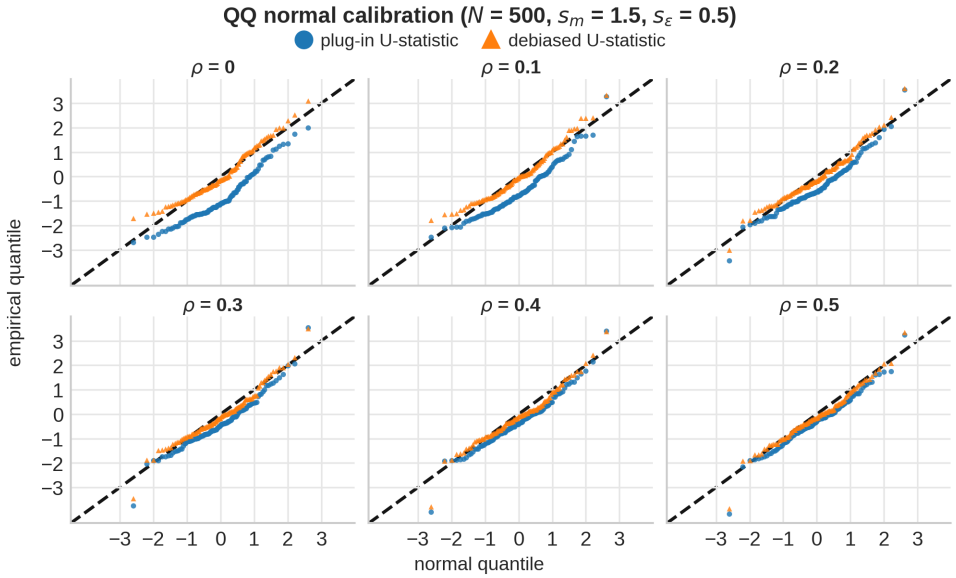


Figure 14: QQ calibration at $n = 500$, $s_m = 1.5$, and $s_\epsilon = 0.5$.

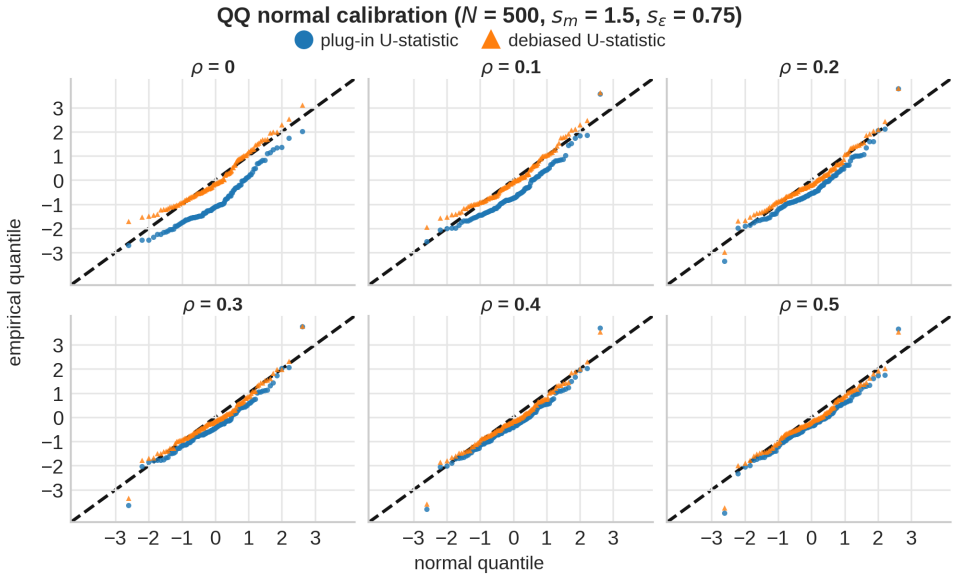


Figure 15: QQ calibration at $n = 500$, $s_m = 1.5$, and $s_\epsilon = 0.75$.

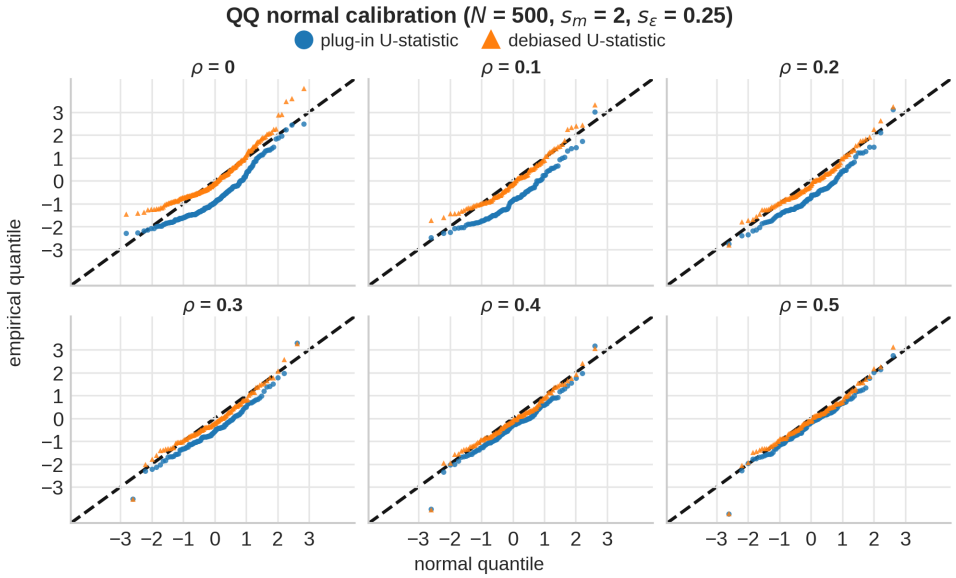


Figure 16: QQ calibration at $n = 500$, $s_m = 2$, and $s_\epsilon = 0.25$.

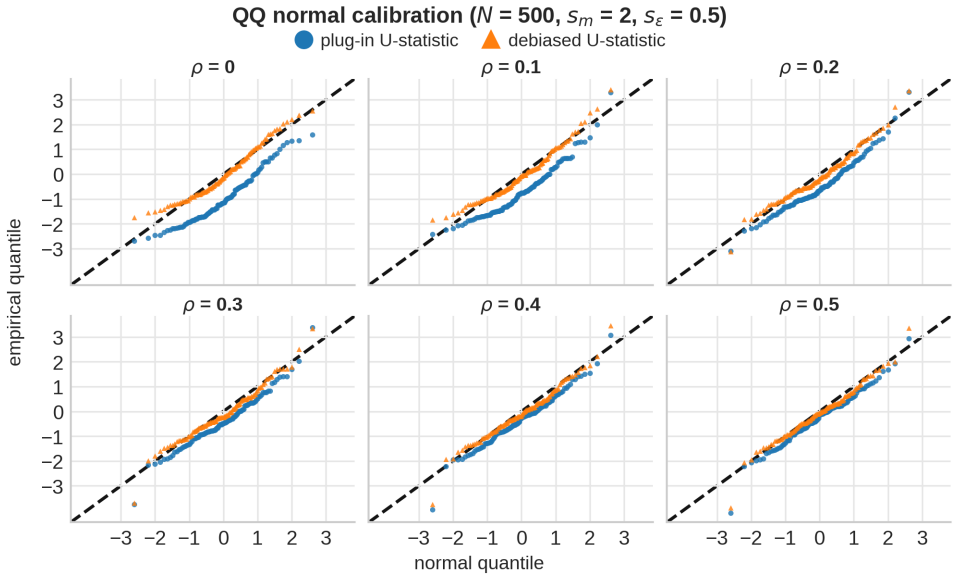


Figure 17: QQ calibration at $n = 500$, $s_m = 2$, and $s_\epsilon = 0.5$.

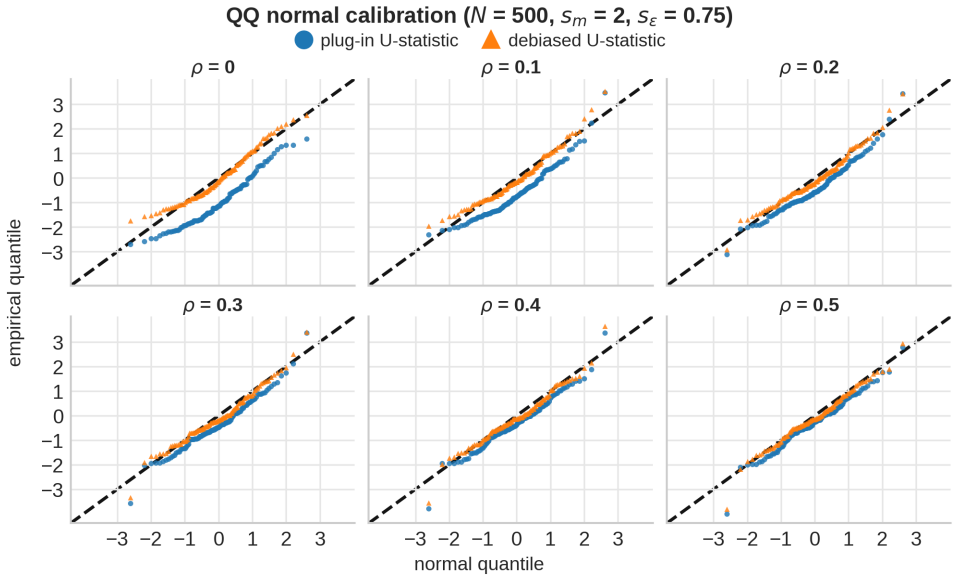


Figure 18: QQ calibration at $n = 500$, $s_m = 2$, and $s_\epsilon = 0.75$.

J.7 Estimation error, coverage, and bandwidth sensitivity

In this subsection we present a more thorough empirical estimation of the dependence of the quality of plug-in and one-step confidence sets in terms of the error of the estimate of Ψ and the choices of kernel bandwidths of $\mathcal{H}_k, \mathcal{H}_l$.

Protocol. We use the random Fourier model of Appendix J.1, with $s_m = 1$, $s_\epsilon = 0.5$, $n \in \{250, 500, 1000\}$, and $\rho \in \{0, 0.1, \ldots, 0.5\}$. Each setting has 100 paired repetitions, with new generating functions and data in each repetition. The plug-in and debiased methods share the data, five cross-fitting folds, mean fits, and kernel bandwidths. We fit the nuisance functions $\hat{m}$ and $\hat{v}$ by kernel ridge regression (vector-valued where appropriate).

Measured quantities. For $a \in \{\mathrm{P}, \mathrm{D}\}$ (plug-in, debiased), let $\widehat{D}_{a,i}$ be the cross-fitted summand of Appendix F, with the EIF correction omitted for $a = \mathrm{P}$. Define

$$\check{\Psi}_{n,a} = \frac{1}{n}\sum_i \widehat{D}_{a,i}, \qquad G_{a,ij} = \langle \widehat{D}_{a,i}, \widehat{D}_{a,j} \rangle_{\mathcal{H}_{kl}}, \qquad U_a = \frac{1}{n(n-1)}\sum_{i \neq j} G_{a,ij}.$$

Writing Ψ_0 for the population operator at each repetition's generating law and selected bandwidths, and $\theta = \|\Psi_0\|^2_{\mathcal{H}_{kl}}$, we report the averages of

$$\|\check{\Psi}_{n,a} - \Psi_0\|^2_{\mathcal{H}_{kl}} \quad \text{(squared error of the operator estimate)},$$
$$(U_a - \theta)^2 \quad \text{(squared error of the HSIC estimate)}.$$

Population norms and inner products use analytic Gaussian expectations over the noise and numerical integration over X. Coverage is the proportion of repetitions in which the nominal 95% interval based on the reverse triangle inequality (RTI) of Appendix G contains $\sqrt{\theta}$. The interval is centered at $\|\check{\Psi}_{n,a}\|_{\mathcal{H}_{kl}}$ and uses 1,000 within-fold bootstrap resamples of the fitted summands. Reported 95% Monte Carlo intervals use 10,000 paired percentile-bootstrap resamples of the repetitions for MSEs and relative gains, and Wilson intervals for coverage.

Median bandwidths. Table 7 reports MSE and coverage at the median-selected bandwidths. The MSE of the debiased operator estimate is approximately 2%–22% higher across the 18 settings. The MSE of the HSIC U-statistic decreases after debiasing at $\rho = 0$ for all sample sizes, and at several positive values of ρ for $n = 500$ and $n = 1000$. At $\rho = 0$, coverage of the debiased RTI interval is 93%, 96%, and 92% for $n = 250, 500, 1000$, respectively; plug-in coverage is 96%, 100%, and 99%. For $\rho \geq 0.2$, coverage is 99%–100% for the debiased RTI intervals and 100% for the plug-in RTI intervals.

Bandwidth sensitivity. Figure 19 fixes $n = 500$ and multiplies the median-selected covariate and residual length scales by $c_X, c_\epsilon \in \{0.1, 0.5, 1, 2, 5\}$. For the Gaussian kernel $\exp(-\|z - z'\|^2/\alpha)$, a multiplier c replaces α by $c^2\alpha$. Mean fits are shared across the grid; the derivative nuisance is retuned when the residual bandwidth changes. For each MSE, the plotted relative reduction is

$$G(c_X, c_\epsilon) = 100\left\{1 - \frac{\widehat{\mathrm{MSE}}_{\mathrm{D}}(c_X, c_\epsilon)}{\widehat{\mathrm{MSE}}_{\mathrm{P}}(c_X, c_\epsilon)}\right\},$$

using the population target for that bandwidth pair. At $c_\epsilon = 0.1$, debiasing increases both MSEs throughout the grid; at $c_\epsilon = 5$, all point estimates show a reduction. Results at intermediate scales vary with ρ and the covariate bandwidth.

Table 7: MSE of the operator estimate and HSIC U-statistic, and coverage of nominal 95% RTI intervals for $\sqrt{\text{HSIC}}$, at median bandwidths, $s_m = 1$, and $s_\epsilon = 0.5$. Entries are estimates with 95% Monte Carlo intervals from 100 paired repetitions.

	Operator MSE ($\times 10^4$)		HSIC U-statistic MSE ($\times 10^7$)		RTI coverage (%)	
n	Plug-in	Debiased	Plug-in	Debiased	Plug-in	Debiased
			(a) $\rho = 0$			
250	9.92 [8.91, 11.07]	10.49 [9.36, 11.79]	4.19 [3.16, 5.51]	3.75 [2.01, 6.00]	96.00 [90.16, 98.43]	93.00 [86.25, 96.57]
500	4.55 [4.16, 4.95]	4.66 [4.26, 5.08]	0.78 [0.65, 0.91]	0.43 [0.30, 0.57]	100.00 [96.30, 100.00]	96.00 [90.16, 98.43]
1000	2.27 [2.04, 2.51]	2.32 [2.09, 2.56]	0.24 [0.20, 0.28]	0.15 [0.10, 0.19]	99.00 [94.55, 99.82]	92.00 [85.00, 95.89]
			(b) $\rho = 0.1$			
250	10.27 [9.24, 11.42]	10.89 [9.72, 12.22]	5.43 [4.09, 7.12]	5.48 [3.32, 8.16]	99.00 [94.55, 99.82]	95.00 [88.82, 97.85]
500	4.67 [4.29, 5.07]	4.83 [4.42, 5.25]	1.25 [1.00, 1.52]	0.97 [0.70, 1.30]	100.00 [96.30, 100.00]	98.00 [93.00, 99.45]
1000	2.31 [2.08, 2.55]	2.36 [2.13, 2.60]	0.52 [0.38, 0.67]	0.40 [0.24, 0.59]	100.00 [96.30, 100.00]	97.00 [91.55, 98.97]
			(c) $\rho = 0.2$			
250	10.53 [9.49, 11.67]	11.35 [10.16, 12.68]	8.36 [6.02, 11.18]	8.90 [5.71, 12.79]	100.00 [96.30, 100.00]	99.00 [94.55, 99.82]
500	4.74 [4.36, 5.11]	5.03 [4.63, 5.46]	2.40 [1.76, 3.11]	2.19 [1.48, 3.00]	100.00 [96.30, 100.00]	99.00 [94.55, 99.82]
1000	2.35 [2.12, 2.59]	2.45 [2.22, 2.71]	1.12 [0.78, 1.51]	1.00 [0.62, 1.47]	100.00 [96.30, 100.00]	99.00 [94.55, 99.82]
			(d) $\rho = 0.3$			
250	10.86 [9.84, 11.93]	12.09 [10.88, 13.41]	11.30 [7.99, 15.30]	13.10 [8.69, 18.37]	100.00 [96.30, 100.00]	99.00 [94.55, 99.82]
500	4.82 [4.46, 5.19]	5.35 [4.95, 5.77]	3.88 [2.72, 5.26]	3.80 [2.52, 5.28]	100.00 [96.30, 100.00]	100.00 [96.30, 100.00]
1000	2.40 [2.18, 2.63]	2.59 [2.35, 2.84]	1.88 [1.28, 2.59]	1.74 [1.11, 2.49]	100.00 [96.30, 100.00]	100.00 [96.30, 100.00]
			(e) $\rho = 0.4$			
250	11.39 [10.36, 12.49]	13.08 [11.74, 14.54]	15.19 [10.38, 21.18]	18.71 [12.05, 26.90]	100.00 [96.30, 100.00]	99.00 [94.55, 99.82]
500	5.10 [4.72, 5.51]	5.78 [5.36, 6.21]	5.67 [3.91, 7.86]	5.76 [3.87, 8.04]	100.00 [96.30, 100.00]	100.00 [96.30, 100.00]
1000	2.45 [2.23, 2.69]	2.77 [2.52, 3.04]	2.69 [1.80, 3.75]	2.67 [1.68, 3.91]	100.00 [96.30, 100.00]	100.00 [96.30, 100.00]
			(f) $\rho = 0.5$			
250	11.55 [10.59, 12.56]	14.03 [12.67, 15.49]	18.41 [12.31, 26.70]	22.62 [14.74, 33.18]	100.00 [96.30, 100.00]	100.00 [96.30, 100.00]
500	5.29 [4.89, 5.72]	6.39 [5.92, 6.91]	7.56 [5.14, 10.66]	7.89 [5.36, 11.08]	100.00 [96.30, 100.00]	100.00 [96.30, 100.00]
1000	2.57 [2.35, 2.81]	3.13 [2.87, 3.43]	3.63 [2.41, 5.09]	3.78 [2.34, 5.63]	100.00 [96.30, 100.00]	100.00 [96.30, 100.00]

(a) $\rho = 0$

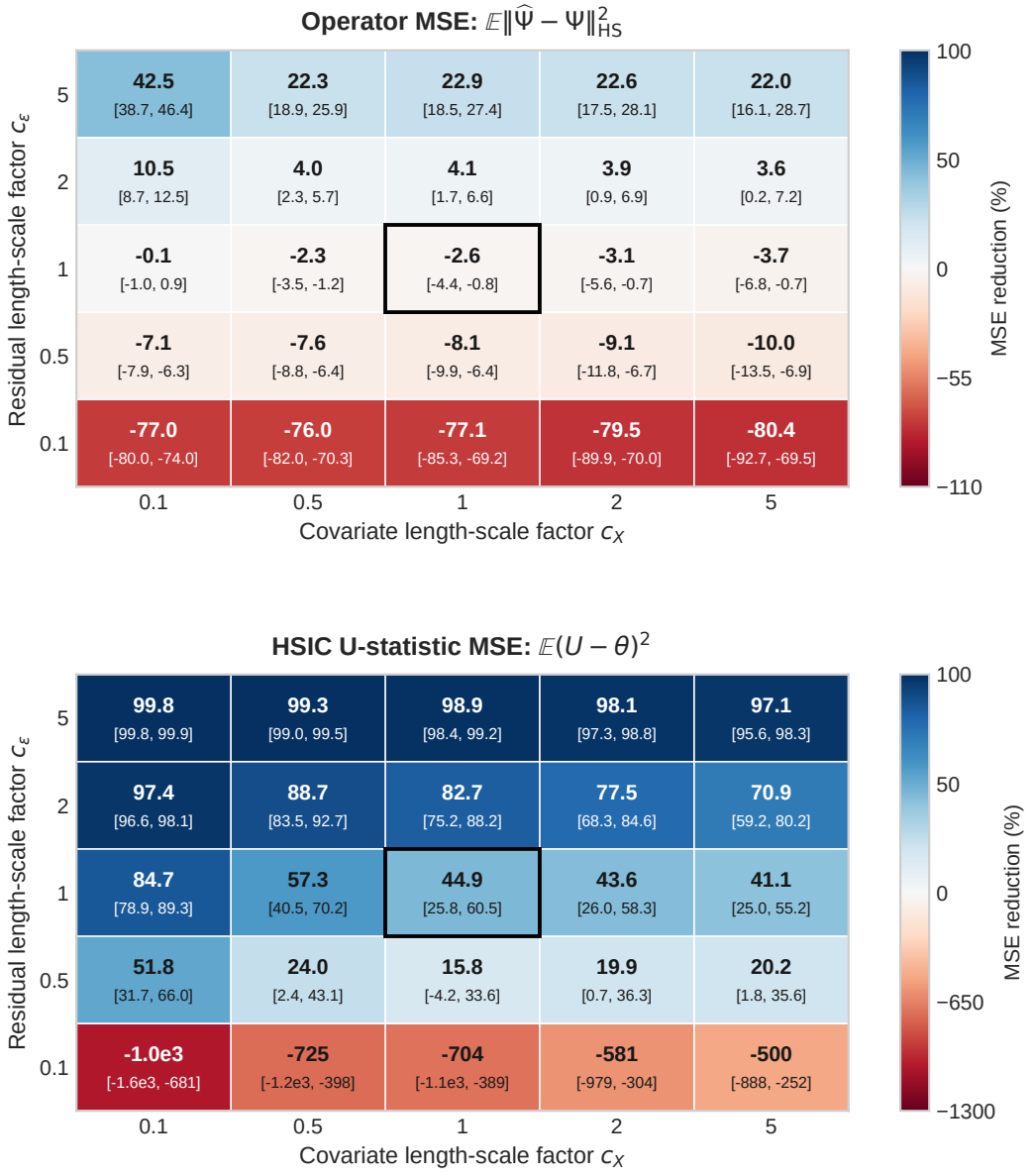
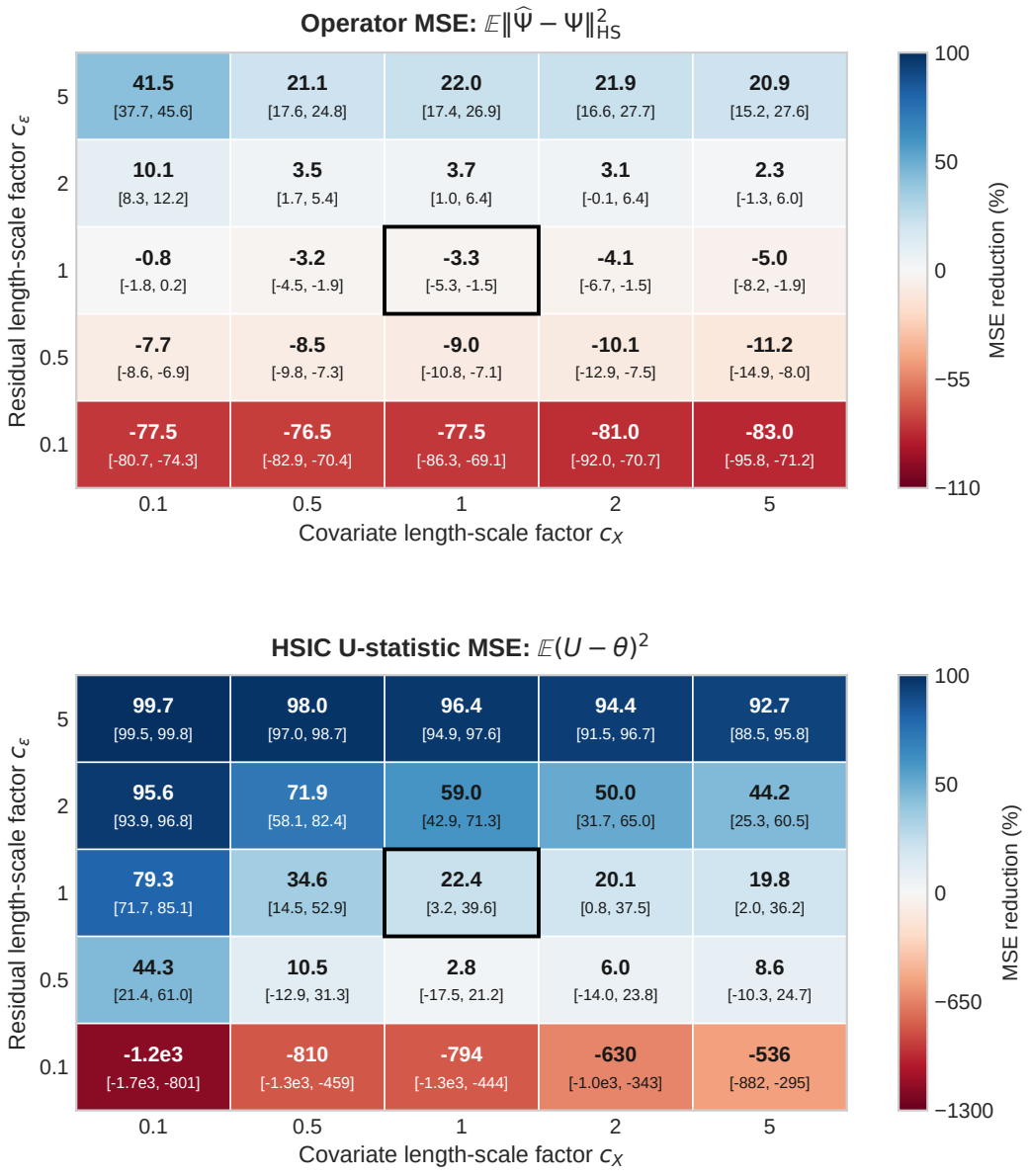


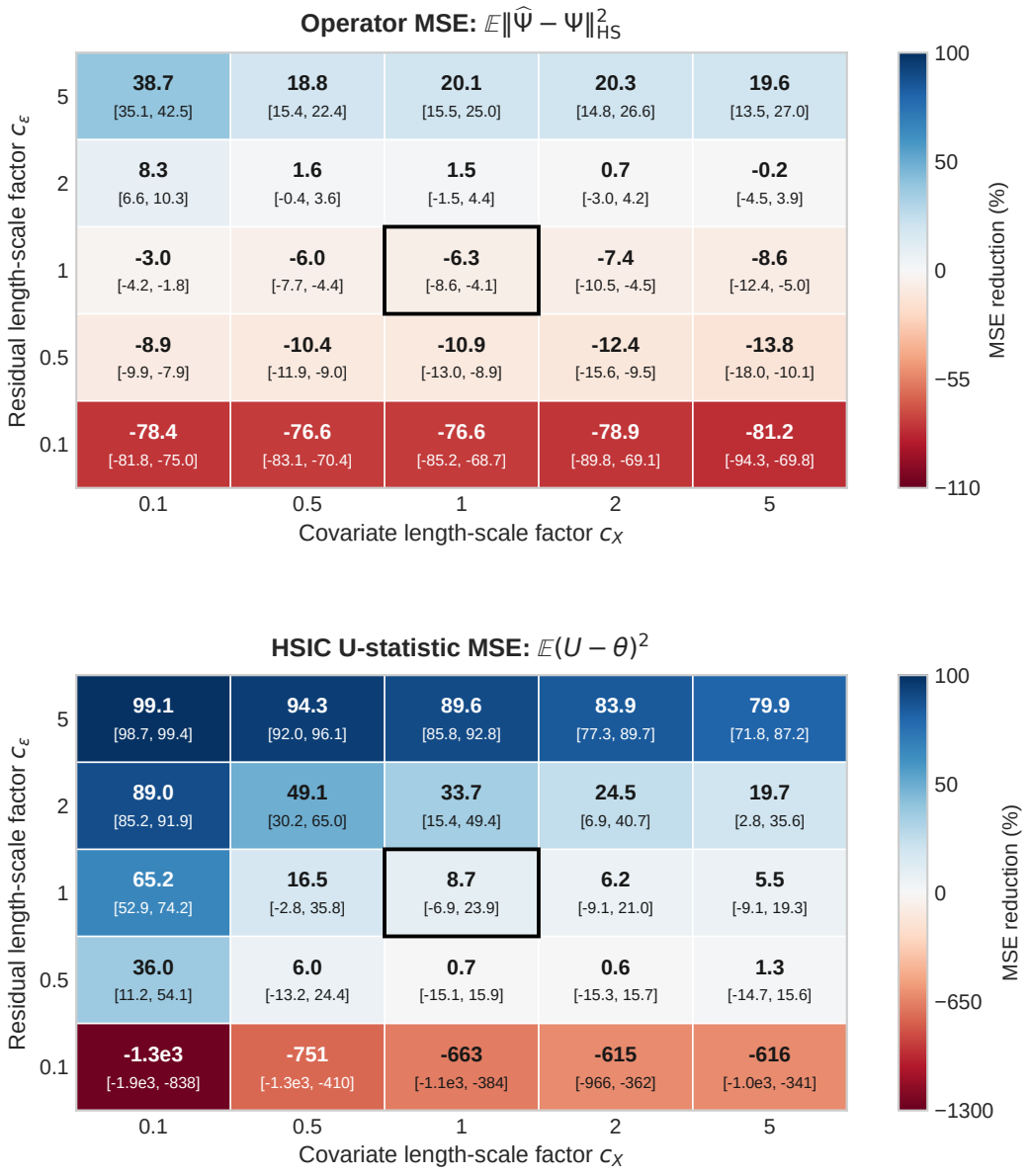
Figure 19: Relative reduction in MSE from debiasing at $n = 500$, $s_m = 1$, and $s_\epsilon = 0.5$, for operator estimation (top) and HSIC U-statistics (bottom). Cells show $G(c_X, c_\epsilon)$ with 95% Monte Carlo intervals. Blue indicates a reduction; red indicates an increase. The outlined cell uses median bandwidths.

(b) $\rho = 0.1$

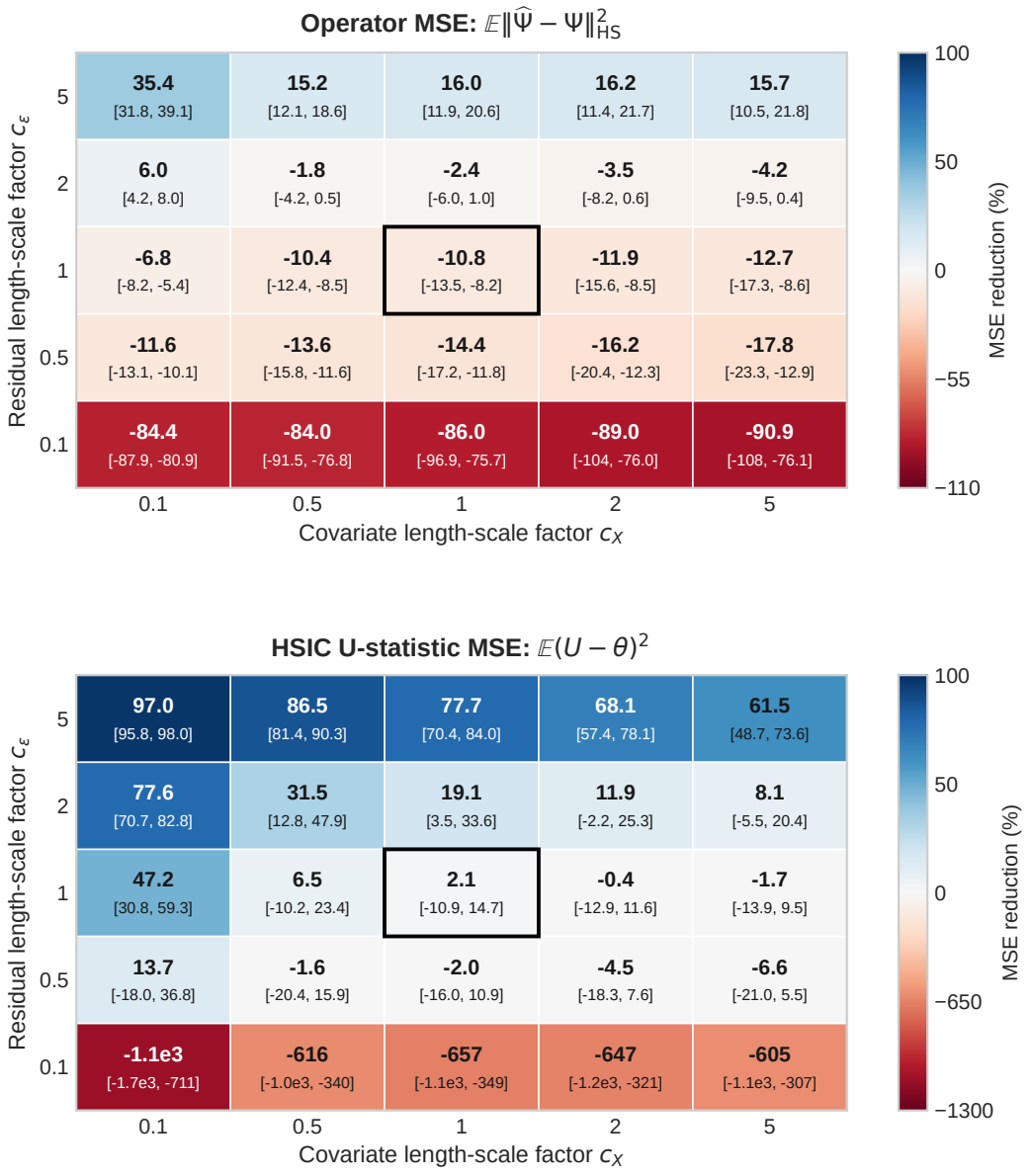

Operator MSE: $\mathbb{E}\|\widehat{\Psi} - \Psi\|^2_{\mathrm{HS}}$
Residual length-scale factor c_ϵ
5
2
1
0.5
0.1
41.5 [37.7, 45.6]
21.1 [17.6, 24.8]
22.0 [17.4, 26.9]
21.9 [16.6, 27.7]
20.9 [15.2, 27.6]
10.1 [8.3, 12.2]
3.5 [1.7, 5.4]
3.7 [1.0, 6.4]
3.1 [-0.1, 6.4]
2.3 [-1.3, 6.0]
-0.8 [-1.8, 0.2]
-3.2 [-4.5, -1.9]
-3.3 [-5.3, -1.5]
-4.1 [-6.7, -1.5]
-5.0 [-8.2, -1.9]
-7.7 [-8.6, -6.9]
-8.5 [-9.8, -7.3]
-9.0 [-10.8, -7.1]
-10.1 [-12.9, -7.5]
-11.2 [-14.9, -8.0]
-77.5 [-80.7, -74.3]
-76.5 [-82.9, -70.4]
-77.5 [-86.3, -69.1]
-81.0 [-92.0, -70.7]
-83.0 [-95.8, -71.2]
0.1
0.5
1
2
5
Covariate length-scale factor c_X
100
50
0
−55
−110
MSE reduction (%)

HSIC U-statistic MSE: $\mathbb{E}(U - \theta)^2$
Residual length-scale factor c_ϵ
5
2
1
0.5
0.1
99.7 [99.5, 99.8]
98.0 [97.0, 98.7]
96.4 [94.9, 97.6]
94.4 [91.5, 96.7]
92.7 [88.5, 95.8]
95.6 [93.9, 96.8]
71.9 [58.1, 82.4]
59.0 [42.9, 71.3]
50.0 [31.7, 65.0]
44.2 [25.3, 60.5]
79.3 [71.7, 85.1]
34.6 [14.5, 52.9]
22.4 [3.2, 39.6]
20.1 [0.8, 37.5]
19.8 [2.0, 36.2]
44.3 [21.4, 61.0]
10.5 [-12.9, 31.3]
2.8 [-17.5, 21.2]
6.0 [-14.0, 23.8]
8.6 [-10.3, 24.7]
-1.2e3 [-1.7e3, -801]
-810 [-1.3e3, -459]
-794 [-1.3e3, -444]
-630 [-1.0e3, -343]
-536 [-882, -295]
0.1
0.5
1
2
5
Covariate length-scale factor c_X
100
50
0
−650
−1300
MSE reduction (%)

(c) $\rho = 0.2$

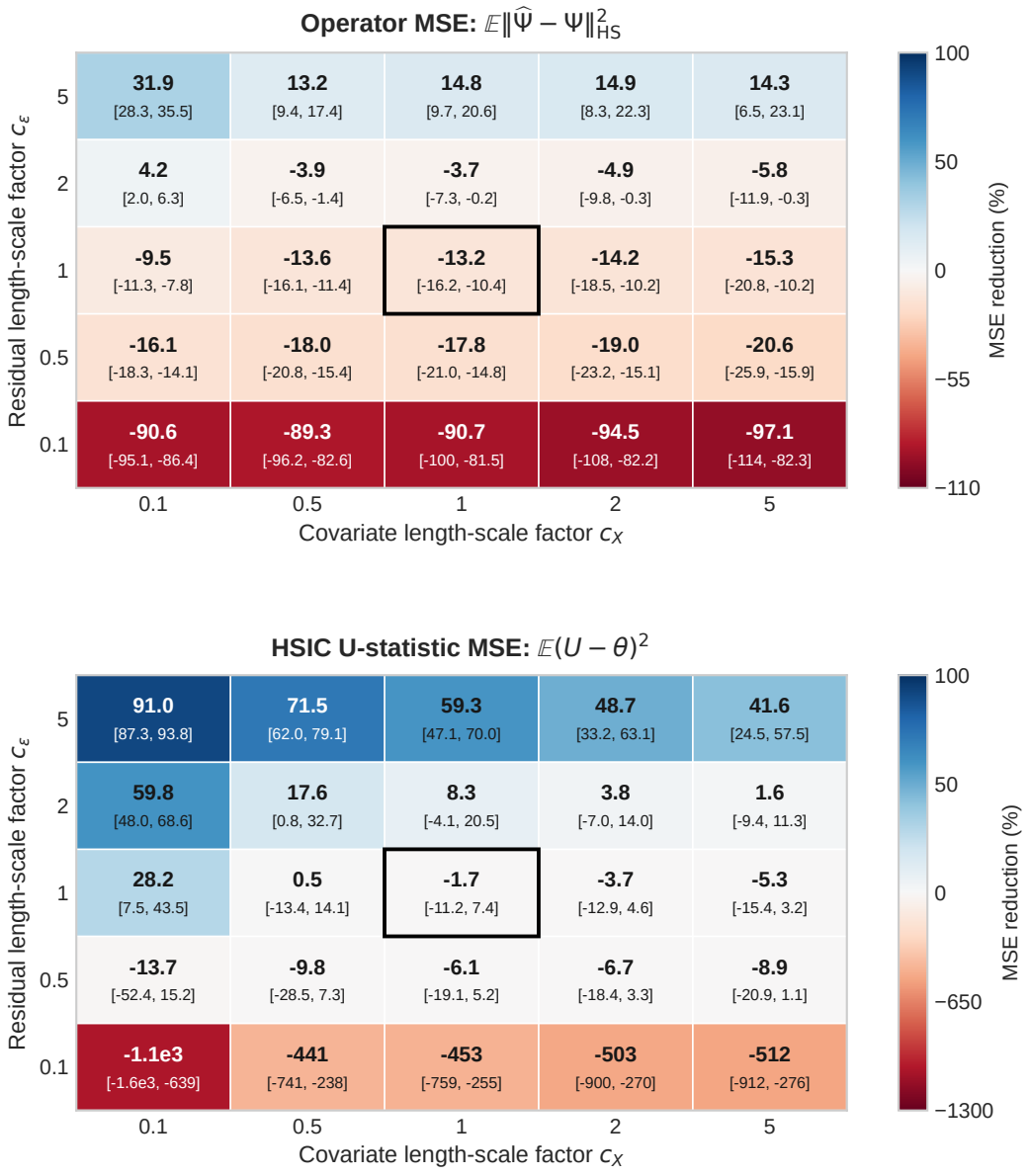

Operator MSE: $\mathbb{E}\|\widehat{\Psi} - \Psi\|^2_{\mathrm{HS}}$
38.7 [35.1, 42.5]
18.8 [15.4, 22.4]
20.1 [15.5, 25.0]
20.3 [14.8, 26.6]
19.6 [13.5, 27.0]
8.3 [6.6, 10.3]
1.6 [-0.4, 3.6]
1.5 [-1.5, 4.4]
0.7 [-3.0, 4.2]
-0.2 [-4.5, 3.9]
-3.0 [-4.2, -1.8]
-6.0 [-7.7, -4.4]
-6.3 [-8.6, -4.1]
-7.4 [-10.5, -4.5]
-8.6 [-12.4, -5.0]
-8.9 [-9.9, -7.9]
-10.4 [-11.9, -9.0]
-10.9 [-13.0, -8.9]
-12.4 [-15.6, -9.5]
-13.8 [-18.0, -10.1]
-78.4 [-81.8, -75.0]
-76.6 [-83.1, -70.4]
-76.6 [-85.2, -68.7]
-78.9 [-89.8, -69.1]
-81.2 [-94.3, -69.8]
Residual length-scale factor c_ε
5
2
1
0.5
0.1
0.1
0.5
1
2
5
Covariate length-scale factor c_X
100
50
0
−55
−110
MSE reduction (%)
HSIC U-statistic MSE: $\mathbb{E}(U - \theta)^2$
99.1 [98.7, 99.4]
94.3 [92.0, 96.1]
89.6 [85.8, 92.8]
83.9 [77.3, 89.7]
79.9 [71.8, 87.2]
89.0 [85.2, 91.9]
49.1 [30.2, 65.0]
33.7 [15.4, 49.4]
24.5 [6.9, 40.7]
19.7 [2.8, 35.6]
65.2 [52.9, 74.2]
16.5 [-2.8, 35.8]
8.7 [-6.9, 23.9]
6.2 [-9.1, 21.0]
5.5 [-9.1, 19.3]
36.0 [11.2, 54.1]
6.0 [-13.2, 24.4]
0.7 [-15.1, 15.9]
0.6 [-15.3, 15.7]
1.3 [-14.7, 15.6]
-1.3e3 [-1.9e3, -838]
-751 [-1.3e3, -410]
-663 [-1.1e3, -384]
-615 [-966, -362]
-616 [-1.0e3, -341]
Residual length-scale factor c_ε
5
2
1
0.5
0.1
0.1
0.5
1
2
5
Covariate length-scale factor c_X
100
50
0
−650
−1300
MSE reduction (%)

(d) $\rho = 0.3$


Operator MSE: $\mathbb{E}\|\widehat{\Psi} - \Psi\|^2_{\mathrm{HS}}$
35.4 [31.8, 39.1]
15.2 [12.1, 18.6]
16.0 [11.9, 20.6]
16.2 [11.4, 21.7]
15.7 [10.5, 21.8]
6.0 [4.2, 8.0]
-1.8 [-4.2, 0.5]
-2.4 [-6.0, 1.0]
-3.5 [-8.2, 0.6]
-4.2 [-9.5, 0.4]
-6.8 [-8.2, -5.4]
-10.4 [-12.4, -8.5]
-10.8 [-13.5, -8.2]
-11.9 [-15.6, -8.5]
-12.7 [-17.3, -8.6]
-11.6 [-13.1, -10.1]
-13.6 [-15.8, -11.6]
-14.4 [-17.2, -11.8]
-16.2 [-20.4, -12.3]
-17.8 [-23.3, -12.9]
-84.4 [-87.9, -80.9]
-84.0 [-91.5, -76.8]
-86.0 [-96.9, -75.7]
-89.0 [-104, -76.0]
-90.9 [-108, -76.1]
Residual length-scale factor c_ϵ
5
2
1
0.5
0.1
0.1
0.5
1
2
5
Covariate length-scale factor c_X
100
50
0
−55
−110
MSE reduction (%)

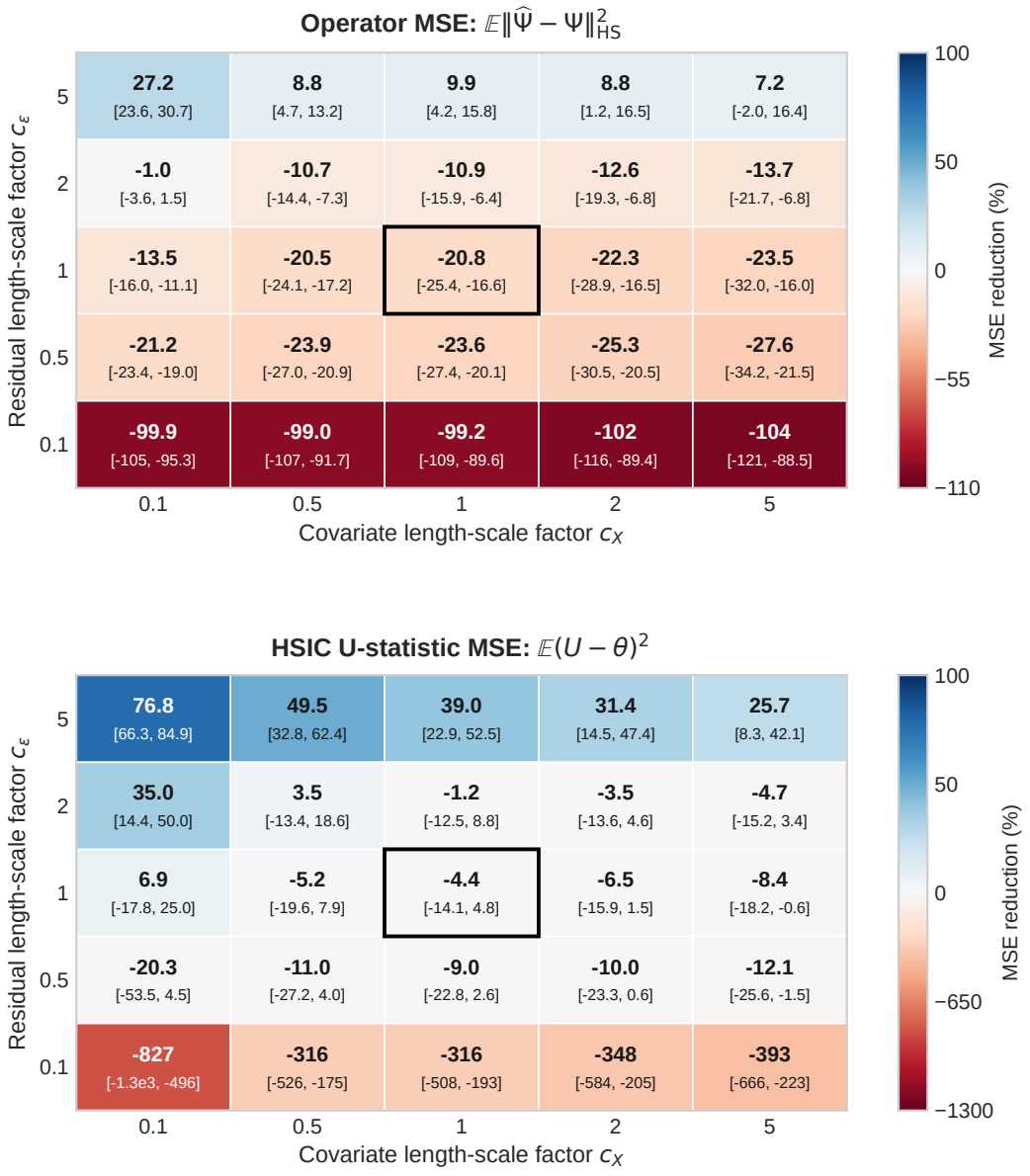
HSIC U-statistic MSE: $\mathbb{E}(U - \theta)^2$
97.0 [95.8, 98.0]
86.5 [81.4, 90.3]
77.7 [70.4, 84.0]
68.1 [57.4, 78.1]
61.5 [48.7, 73.6]
77.6 [70.7, 82.8]
31.5 [12.8, 47.9]
19.1 [3.5, 33.6]
11.9 [-2.2, 25.3]
8.1 [-5.5, 20.4]
47.2 [30.8, 59.3]
6.5 [-10.2, 23.4]
2.1 [-10.9, 14.7]
-0.4 [-12.9, 11.6]
-1.7 [-13.9, 9.5]
13.7 [-18.0, 36.8]
-1.6 [-20.4, 15.9]
-2.0 [-16.0, 10.9]
-4.5 [-18.3, 7.6]
-6.6 [-21.0, 5.5]
-1.1e3 [-1.7e3, -711]
-616 [-1.0e3, -340]
-657 [-1.1e3, -349]
-647 [-1.2e3, -321]
-605 [-1.1e3, -307]
Residual length-scale factor c_ϵ
5
2
1
0.5
0.1
0.1
0.5
1
2
5
Covariate length-scale factor c_X
100
50
0
−650
−1300
MSE reduction (%)

(e) $\rho = 0.4$


Operator MSE: $\mathbb{E}\|\widehat{\Psi} - \Psi\|^2_{\mathrm{HS}}$
31.9 [28.3, 35.5]
13.2 [9.4, 17.4]
14.8 [9.7, 20.6]
14.9 [8.3, 22.3]
14.3 [6.5, 23.1]
4.2 [2.0, 6.3]
-3.9 [-6.5, -1.4]
-3.7 [-7.3, -0.2]
-4.9 [-9.8, -0.3]
-5.8 [-11.9, -0.3]
-9.5 [-11.3, -7.8]
-13.6 [-16.1, -11.4]
-13.2 [-16.2, -10.4]
-14.2 [-18.5, -10.2]
-15.3 [-20.8, -10.2]
-16.1 [-18.3, -14.1]
-18.0 [-20.8, -15.4]
-17.8 [-21.0, -14.8]
-19.0 [-23.2, -15.1]
-20.6 [-25.9, -15.9]
-90.6 [-95.1, -86.4]
-89.3 [-96.2, -82.6]
-90.7 [-100, -81.5]
-94.5 [-108, -82.2]
-97.1 [-114, -82.3]
Residual length-scale factor c_ϵ
5
2
1
0.5
0.1
0.1
0.5
1
2
5
Covariate length-scale factor c_X
100
50
0
−55
−110
MSE reduction (%)

HSIC U-statistic MSE: $\mathbb{E}(U - \theta)^2$
91.0 [87.3, 93.8]
71.5 [62.0, 79.1]
59.3 [47.1, 70.0]
48.7 [33.2, 63.1]
41.6 [24.5, 57.5]
59.8 [48.0, 68.6]
17.6 [0.8, 32.7]
8.3 [-4.1, 20.5]
3.8 [-7.0, 14.0]
1.6 [-9.4, 11.3]
28.2 [7.5, 43.5]
0.5 [-13.4, 14.1]
-1.7 [-11.2, 7.4]
-3.7 [-12.9, 4.6]
-5.3 [-15.4, 3.2]
-13.7 [-52.4, 15.2]
-9.8 [-28.5, 7.3]
-6.1 [-19.1, 5.2]
-6.7 [-18.4, 3.3]
-8.9 [-20.9, 1.1]
-1.1e3 [-1.6e3, -639]
-441 [-741, -238]
-453 [-759, -255]
-503 [-900, -270]
-512 [-912, -276]
Residual length-scale factor c_ϵ
5
2
1
0.5
0.1
0.1
0.5
1
2
5
Covariate length-scale factor c_X
100
50
0
−650
−1300
MSE reduction (%)

(f) $\rho = 0.5$


Operator MSE: $\mathbb{E}\|\widehat{\Psi} - \Psi\|^2_{\mathrm{HS}}$
27.2 [23.6, 30.7]
8.8 [4.7, 13.2]
9.9 [4.2, 15.8]
8.8 [1.2, 16.5]
7.2 [-2.0, 16.4]
-1.0 [-3.6, 1.5]
-10.7 [-14.4, -7.3]
-10.9 [-15.9, -6.4]
-12.6 [-19.3, -6.8]
-13.7 [-21.7, -6.8]
-13.5 [-16.0, -11.1]
-20.5 [-24.1, -17.2]
-20.8 [-25.4, -16.6]
-22.3 [-28.9, -16.5]
-23.5 [-32.0, -16.0]
-21.2 [-23.4, -19.0]
-23.9 [-27.0, -20.9]
-23.6 [-27.4, -20.1]
-25.3 [-30.5, -20.5]
-27.6 [-34.2, -21.5]
-99.9 [-105, -95.3]
-99.0 [-107, -91.7]
-99.2 [-109, -89.6]
-102 [-116, -89.4]
-104 [-121, -88.5]
Residual length-scale factor c_ϵ
5
2
1
0.5
0.1
0.1
0.5
1
2
5
Covariate length-scale factor c_X
100
50
0
−55
−110
MSE reduction (%)

HSIC U-statistic MSE: $\mathbb{E}(U - \theta)^2$
76.8 [66.3, 84.9]
49.5 [32.8, 62.4]
39.0 [22.9, 52.5]
31.4 [14.5, 47.4]
25.7 [8.3, 42.1]
35.0 [14.4, 50.0]
3.5 [-13.4, 18.6]
-1.2 [-12.5, 8.8]
-3.5 [-13.6, 4.6]
-4.7 [-15.2, 3.4]
6.9 [-17.8, 25.0]
-5.2 [-19.6, 7.9]
-4.4 [-14.1, 4.8]
-6.5 [-15.9, 1.5]
-8.4 [-18.2, -0.6]
-20.3 [-53.5, 4.5]
-11.0 [-27.2, 4.0]
-9.0 [-22.8, 2.6]
-10.0 [-23.3, 0.6]
-12.1 [-25.6, -1.5]
-827 [-1.3e3, -496]
-316 [-526, -175]
-316 [-508, -193]
-348 [-584, -205]
-393 [-666, -223]
Residual length-scale factor c_ϵ
5
2
1
0.5
0.1
0.1
0.5
1
2
5
Covariate length-scale factor c_X
100
50
0
−650
−1300
MSE reduction (%)

K Comparison with classical residual diagnostics

We compare our method with classical tests of residual variability and a quantile-regression diagnostic in the causal-discovery experiment of Section 4 and appendix J.3. To ensure a fair comparison, all methods use the same nonparametric mean regression to form their residuals.

We describe the tests for the forward direction, where X is the predictor, Y is the response, and $W = X$. As in Section 2, the population regression is $m_0(x) = \mathbb{E}_0[Y \mid X = x]$ and the population residual is $\xi_0(X, Y) = Y - m_0(X)$, with $\mathbb{E}_0$ denoting expectation under the true law. For the observed pairs (x_i, y_i), $i = 1, \ldots, n$, where n is the sample size, we use estimated residuals $\widehat{\xi}_i = y_i - \widehat{m}(x_i)$. Here $\widehat{m}$ is the regression estimate trained on the appropriate training fold or split, which excludes the observation being evaluated. The reverse-direction tests interchange the roles of X and Y.

Breusch–Pagan and White-type tests. To apply these tests in our setting, we replace their usual parametric mean regression with the nonparametric fit above. These tests ask whether the squared residuals are linearly associated with chosen features of the predictor [Breusch and Pagan, 1979, White, 1980]. We therefore regress $\widehat{\xi}_i^2$ on an intercept and a finite-dimensional predictor feature map φ_k. For a predictor value x, write $u(x) = (x - \bar{x})/s_x$, where $\bar{x}$ and s_x are the predictor sample mean and standard deviation. We consider polynomial features of degree d,

$$\varphi_k(x) = \big(u(x), \ldots, u(x)^d\big)^\top, \qquad d \in \{1, 2, 3\},$$

and Fourier features with J frequencies,

$$\varphi_k(x) = \big(\sin u(x), \cos u(x), \ldots, \sin(Ju(x)), \cos(Ju(x))\big)^\top, \qquad J \in \{1, 2, 4, 8\}.$$

The linear and quadratic specifications give the Breusch–Pagan and one-dimensional White tests, respectively. We use the studentized version [Koenker, 1981], extending the same construction to the other features.

For fixed predictor features, these tests are closely related to our cross-covariance target. For a real residual value r, take the residual feature to be $\varphi_l(r) = r^2$. The target then becomes $\mathrm{Cov}_0(\varphi_k(X), \xi_0(X, Y)^2)$, where Cov_0 denotes covariance under the true law: a finite-dimensional instance of Ψ_0, provided the required moments exist. The induced kernel is $l(r, r') = r^2(r')^2$ for real residual values r, r', so it falls outside our bounded-kernel assumption, Assumption 2, on an unbounded domain. Since the derivative of this residual feature is $\varphi_l'(r) = 2r$ and the population residual has conditional mean zero,

$$\mathbb{E}_0[\varphi_l'(\xi_0(X, Y)) \mid X] = 2\mathbb{E}_0[\xi_0(X, Y) \mid X] = 0.$$

Thus the first-order effect of estimating the mean, and hence our first-order correction, vanishes for this feature; higher-order estimation errors can still remain. For this variance-only feature, the main difference is how the estimated dependence is used: our method constructs a confidence interval for the cross-covariance norm, whereas BP/White scale the feature associations by their estimated uncertainty before testing whether they are zero.

Levene and Brown–Forsythe tests. These tests compare residual variability across groups [Brown and Forsythe, 1974]. To apply them to a continuous predictor, we sort the observed values of X and divide them into B bins containing as nearly equal numbers of observations as possible, with $B \in \{2, 3, 4, 5, 8, 10\}$. For each residual, we calculate its absolute deviation from the mean residual in its bin for Levene's test, or from the median residual in its bin for Brown–Forsythe's test. A one-way analysis of variance (ANOVA) then tests whether the mean absolute deviations differ across bins.

Quantile-regression diagnostic. Under the null of residual independence, every conditional quantile of $\xi_0(X, Y)$ is constant in X. This suggests fitting quantile regressions to the estimated residuals and testing whether their predictor coefficients vanish [Koenker and Bassett, 1978]. For each quantile level τ in a finite set $\mathcal{T}$, let $\widehat{a}_\tau$ and $\widehat{b}_\tau$ be the fitted intercept and vector of predictor coefficients. The estimated conditional τ-quantile at predictor value x, denoted by $\widehat{Q}_\tau(x)$, is

$$\widehat{Q}_\tau(x) = \widehat{a}_\tau + \varphi_k(x)^\top \widehat{b}_\tau,$$

For each quantile, we use a Wald test of zero predictor coefficients with a covariance estimate consistent under heteroscedasticity. We combine the tests using a Bonferroni correction: multiply the smallest p-value by the number of quantiles, and cap the result at one. We consider the polynomial features above with $d \in \{1, 2, 3\}$ and Fourier features with $J \in \{2, 4\}$, using the quantile grids $\mathcal{T} = \{0.25, 0.5, 0.75\}$ and $\mathcal{T} = \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

Experimental protocol. We evaluated both five-fold cross-fitting and sample splitting, with training fractions $\{0.2, 0.4, 0.6, 0.8\}$, in a preliminary comparison of 100 repetitions. Table 8 reports the configurations retained from that comparison, evaluated on 1,000 fresh repetitions with $n = 250$ and nominal level 5%. All retained configurations use five-fold cross-fitting, and all methods use identical estimated residuals within each repetition and regression direction. Each mean regression is Gaussian kernel ridge regression trained on 200 observations. Its ridge penalty and kernel length scale are chosen by five-fold cross-validation on the training data, minimizing squared-error loss over $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ and $\{0.5, 1, 2, 4, 8\}$, respectively. In the table, ρ is the noise-heterogeneity parameter from Appendix J.3: the forward direction has independent noise at $\rho = 0$ and heterogeneous noise at $\rho = 0.35$.

Results and calibration. Debiased HSIC rejects most often in both reverse-direction alternatives, while most methods have power close to one in the heterogeneous forward direction. Debiased HSIC, BP/White, and Levene reject more often than the nominal 5%, whereas plug-in HSIC and the quantile diagnostic are conservative. Brown–Forsythe's confidence interval contains 5%.

L Witness functions and their interpretation

In this section we describe how the learned estimators of the cross-covariance operator Ψ_0 can be used to interpret the nature of the null violation. We consider Setting 1 of Section 2, with $W = X$ and scalar X and Y. Let $R = \xi_0(X, Y) = Y - m_0(X)$ and $\mathcal{H}_{kl} = \mathcal{H}_k \otimes \mathcal{H}_l$. Under the canonical identification of Hilbert–Schmidt operators with functions in $\mathcal{H}_{kl}$, our cross-covariance operator Ψ_0

Table 8: Rejection probabilities at nominal level 5% in the causal-discovery experiment, using 1,000 fresh repetitions with $n = 250$; brackets give pointwise 95% Clopper–Pearson intervals. Only $\rho = 0$, $X \to Y$ is a null-size experiment; the other columns report power.

	$\rho = 0$		$\rho = 0.35$	
Method	$X \to Y$ (size)	$Y \to X$ (power)	$X \to Y$ (power)	$Y \to X$ (power)
Plug-in HSIC (permutation)	0.010 [0.005, 0.018]	0.126 [0.106, 0.148]	0.994 [0.987, 0.998]	0.324 [0.295, 0.354]
Debiased HSIC (ours)	0.086 [0.069, 0.105]	0.483 [0.452, 0.514]	0.999 [0.994, 1.000]	0.658 [0.628, 0.687]
BP/White: linear	0.071 [0.056, 0.089]	0.062 [0.048, 0.079]	0.984 [0.974, 0.991]	0.349 [0.319, 0.379]
BP/White: Fourier $J = 2$	0.107 [0.089, 0.128]	0.238 [0.212, 0.266]	1.000 [0.996, 1.000]	0.386 [0.356, 0.417]
BP/White: Fourier $J = 4$	0.111 [0.092, 0.132]	0.354 [0.324, 0.385]	1.000 [0.996, 1.000]	0.444 [0.413, 0.475]
Levene: $B = 10$	0.085 [0.068, 0.104]	0.407 [0.376, 0.438]	1.000 [0.996, 1.000]	0.531 [0.500, 0.562]
Brown–Forsythe: $B = 10$	0.045 [0.033, 0.060]	0.184 [0.160, 0.209]	0.998 [0.993, 1.000]	0.291 [0.263, 0.320]
Quantile regression: Fourier $J = 4$	0.006 [0.002, 0.013]	0.226 [0.200, 0.253]	0.348 [0.318, 0.378]	0.288 [0.260, 0.317]

All methods use five-fold cross-fitting. The quantile diagnostic uses $\mathcal{T} = \{0.25, 0.5, 0.75\}$. The Gaussian-kernel HSIC procedures use 399 permutations or bootstrap draws, respectively, and reject when their p-value is at most 0.05; the classical diagnostics use a strict 0.05 threshold.

is the HSIC witness function:

$$w(x,r) = \langle \Psi_0, \varphi_k(x) \otimes \varphi_l(r) \rangle_{\mathcal{H}_{kl}} = \mathbb{E}_0[k(x,X)l(r,R)] - \mathbb{E}_0[k(x,X)]\mathbb{E}_0[l(r,R)].$$

See Jitkrittum et al. [2017, Section 2.1, equation (2)]. Our cross-fitted one-step estimator $\check{\Psi}_n$ (Remark 1 and appendix F) gives the estimated witness, with $\widehat{w}_n(x,r) = \langle \check{\Psi}_n, \varphi_k(x) \otimes \varphi_l(r) \rangle_{\mathcal{H}_{kl}}$. For $\Psi_0 \neq 0$, the optimal unit-norm witness is $w/\|\Psi_0\|_{\mathcal{H}_{kl}}$, and $\mathrm{HSIC}(X,R) = \|\Psi_0\|^2_{\mathcal{H}_{kl}}$. Throughout this section we show unnormalised witnesses.

A positive value at (x,r) indicates a kernel-smoothed excess of joint mass near (x,r) relative to the independence law $P_X \otimes P_R$; a negative value indicates a deficit. The full map therefore shows where the residual distribution depends on the covariate. The top row of Figure 20 compares the population witness, the empirical oracle using the true regression m_0, and our debiased operator estimate using fitted regressions, all for the same configuration and kernels.

For a fixed x, the marginal residual density is

$$p_R(t) = \int p_{R|X=u}(t)\, dP_X(u).$$

With $a(x) = \mathbb{E}_0[k(x,X)] > 0$, the conditional residual density smoothed using k and the witness slice are

$$p_{R;k,x}(t) = \frac{\mathbb{E}_0[k(x,X)p_{R|X}(t \mid X)]}{a(x)}, \qquad w(x,r) = a(x) \int l(r,t)\big[p_{R;k,x}(t) - p_R(t)\big]\, dt.$$

Thus the slice is a smoothed contrast between the kernel-averaged conditional noise law $p_{R;k,x}$ and the marginal noise law p_R. For a locally smooth conditional law and x in the covariate support, a small covariate kernel bandwidth makes $p_{R;k,x}$ close to $p_{R|X=x}$.

The lower panels illustrate this contrast: at $x = -1.5$, the positive centre and negative tails indicate more concentrated noise; at $x = 2.5$ the signs reverse. At $x = 0.5$ the two averaged laws nearly agree and the slice is small. Here $R \mid X = x \sim \mathcal{N}(0, \sigma(x)^2)$, as in the synthetic model of Section 4 and appendix J.1, so dependence arises entirely from heteroskedasticity. Under independence the population witness is zero everywhere.

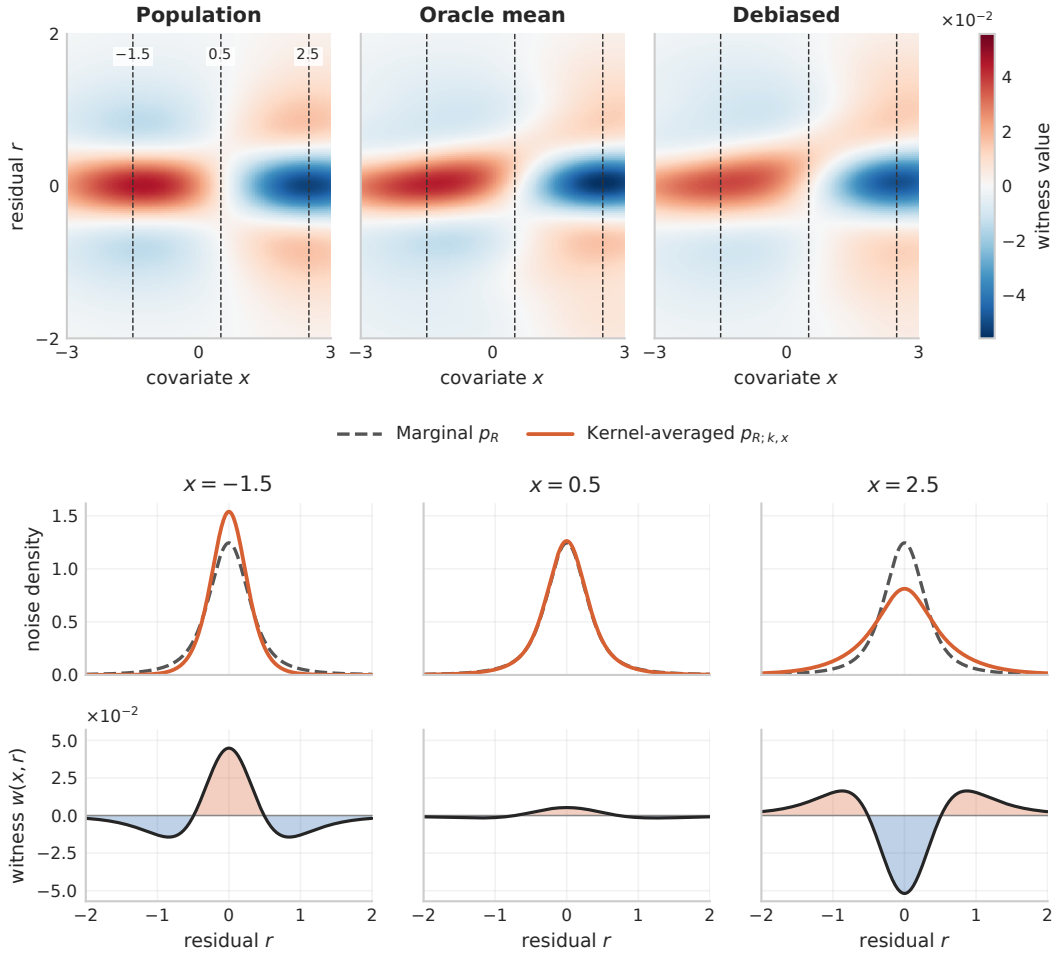


Figure 20: Full witnesses and interpretation of population slices. Top: population, finite-sample oracle-mean, and debiased witnesses on a common colour scale; dashed lines mark the three slice locations. Both estimates use the same $n = 500$ observations and five-fold centering; the oracle uses true residuals, and the debiased estimate uses residuals from the fitted mean regression together with the fitted derivative correction. Middle: the X-independent marginal noise density p_R (grey dashed) and the kernel-averaged conditional density $p_{R;k,x}$ (orange). Bottom: the population slice, with positive and negative values shaded orange and blue. Axes are shared within each row. We fix $\rho = 0.5$, $s_m = 1$, $s_\epsilon = 0.75$, and $X \sim \mathrm{Unif}[-3, 3]$. All panels use the same Gaussian kernels, with bandwidths selected by the median heuristic.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that we develop Hilbert-valued one-step estimators for kernel measures of residual noise heterogeneity, together with tests and confidence intervals. The theoretical and empirical claims are stated under the assumptions used in the analysis, and are proved, referenced, or evaluated empirically.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the nonregular behavior of the scalar norm at the null, state the assumptions under which our asymptotic results hold, and describe the quadratic computational scaling of the kernel computations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Every mathematical result is stated as a precise mathematical statement with assumptions or references. Proofs are provided in the appendix or cited where the result is imported from prior work.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We explain the experimental procedures in the experimental section and the relevant appendix. We also provide a ZIP file with code for reproducing the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide a ZIP file with self-contained code, up to standard Python dependencies, and scripts for reproducing the reported synthetic experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe our data-generating processes, fitting procedures, kernel choices, resampling counts, and compute details in the appendix. The released code exposes the relevant parameters in readable configuration files.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: When reporting individual empirical rejection probabilities or coverage estimates, we accompany them with 95% error bars. Aggregate coverage summaries are labeled as ranges of observed empirical point estimates rather than confidence intervals.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the available computational resources, representative runtime, and quadratic sample-size scaling of the kernel computations in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is methodological, uses synthetic datasets, and does not involve human subjects, sensitive data, or deployed systems.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper presents foundational statistical methodology evaluated on synthetic data, and is not tied to a deployed system or a high-risk application. Potential applications such as causal discovery and residual diagnostics are discussed only at a scientific level.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk models, scraped datasets, or assets with substantial misuse potential; the released assets are code and synthetic data-generation scripts.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use or repackage existing datasets, pretrained models, or other external assets. Existing methods used for comparison or background are cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The provided code is documented and includes scripts and configuration files for reproducing the experiments and generating the synthetic data.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, data annotation, or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does ***not*** impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as part of the core method, experiments, or scientific analysis. Any use was limited to standard wording, editing, or formatting assistance.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.