Consistent Test for Conditional Moment Restriction Models in Reproducing Kernel Hilbert Spaces

Yuhao Li LIYUHAO.ECON@WHU.EDU.CN
Economics and Management School
Wuhan University

Xiaojun Song SXJ@GSM.PKU.EDU.CN
Guanghua School of Management
Peking University

Draft Time: November, 2022. Click Here for Latest Version

Abstract

In this paper, we represent *Integrated Conditional Moment* (ICM) tests in *Reproducing Kernel Hilbert Spaces* (RKHS). There are several advantages of doing so. First, reproducing kernels embody dimension and integral measure, and hence, are effective dimension reduction tools. This phenomenon can be explained by the isometrically isomorphic relationship among infinite dimensional Hilbert spaces. Second, the test statistics, expressed in terms of kernels, have analytic closed forms, making them easy to compute in practice. Third, one can generate kernels easily and massively from existing kernels. Each kernel corresponds to an ICM test, thus, for certain models, one may obtain a more sensitive test than by using conventional ones. We further propose projection-based kernels to eliminate estimation effect, leading to a simple multiplier bootstrap procedure to obtain critical values. A minimum distance estimator is developed as a byproduct. Monte Carlo exercises are performed to examine the finite sample performance of the proposed test, and an empirical application is studied.

JEL Classification: C12; C13; C15; C52

Keywords: Conditional Moment Restriction, Specification Tests, Reproducing Kernel Hilbert Spaces, Multiplier Bootstrap

1. Introduction

In this paper, we study the question of testing conditional moment restriction (CMR) models. Specifically, we develop a new framework for deriving integrated conditional moment (ICM) test statistics. This framework is based on the idea of embedding CMR in a reproducing kernel Hilbert space (RKHS). The test statistic is defined as the maximum moment restriction (MMR) within the unit ball of the RKHS. Furthermore, we show that MMR corresponds to the RKHS norm of a Hilbert space embedding of conditional moments.

We contribute to the literature in the following aspects.

Closed Form Expression. ICM statistics, as its name suggests, is obtained after integrating the nuisance parameter of infinite many unconditional moment restrictions (UMR). This integration often leads to numerical challenges, and only few weighting functions and integral measures are known in the literature to generate closed-form statistics, e.g., Bierens (1982); Escanciano (2006a). In our framework, the MMR is obtained directly from a user-chosen reproducing kernel without integration. Furthermore, we show that the MMR captures all information about the original CMR and it has a closed-form expression which eases practical implementation.

Dimension Reduction. The dimension of conditional variables often poses practical and theoretical challenges when conducting CMR specification tests. Most of the existing ICM tests depend on high-dimensional stochastic processes, e.g., Domínguez and Lobato (2015), and their power performance often drops significantly as the dimension d increases due to the data sparseness. One common solution is to project the original conditional covariates X onto the $\beta^\top X$ for all $\|\beta\|_2 = 1$, see, e.g., Escanciano (2006a); Lavergne and Patilea (2012); Sant'Anna and Song (2019). Here, $\|\cdot\|_2$ denotes the Euclidean norm. However, due to the involvement of infinite directions, projection-based tests are often computational intensive and the powers of these tests are often low (Guo and Zhu, 2017). Reproducing kernels, on the other hand, embody both the dimension and the integral measure. As a result, the estimator of the MMR converges in the RKHS norm in a way that is independent of the dimension. This is an appealing property since testings based on this estimator are less sensitive to the curse of dimensionality.

Massively Generate New Tests. Existing literature has shown that the power of an ICM test is determined by the weighting function, the integral measure, the data generating process (DGP) and the model itself. See, e.g., Escanciano (2009). Thus, an ICM test statistic might be powerful against one model and one DGP, but could be powerless against another model or another DGP. Hence, it is desirable to have as many ICM test statistics as possible. We provide methods to construct new kernels from existing kernels. Since each kernel corresponds to an ICM test statistic, this means that one could generate infinite many new ICM test statistics, all have closed-form expression.

Eliminate Estimation Effect. We propose a projected kernel to cancel the estimation effect. The limiting null distribution, therefore, does not depend on how an estimator is obtained and does not require the estimator to be $\sqrt{n}$-asymptotically linear, with n the sample size. Without the estimation effect, critical values are obtained via a simple and fast multiplier bootstrap procedure, and perhaps most interestingly, the proposed test is capable of applying to certain 'non-standard' estimators who has a slower convergence rate.

New Limit Distribution. We derive the limit distribution of the proposed test statistic under the fixed alternative. This new result provides a framework to obtain a more powerful test by selecting an optimal kernel. Nevertheless, much work is needed to fully achieve this goal.

A Minimum Distance Estimator. We propose a minimum distance estimator based on the MMR. Comparing to existing minimum distance estimators, e.g., Domínguez and Lobato (2004), it has the advantage of less sensitive to the curse of dimensionality.

The rest of the paper is organized as follows. Section 2 presents our main idea of using RKHS framework to develop the test statistic. We discuss the some merits of doing so, as well as the challenges when unknown parameters are replaced by their estimators. Section 3 describes a method for projecting a kernel onto a tangent space of nuisance parameters, so that the modified statistic is free from the estimation effect. We also establish the asymptotic properties of our test in this section. In section 4, we introduce a simple multiplier bootstrap procedure to obtain critical values, and justify its asymptotic validity. Based on the test statistic, we propose a minimum distance estimator, its asymptotic properties are studied in Section 5. Section 6 conducts Monte Carlo experiments. Simulation results indicate that the proposed tests have an accurate finite sample size and well local power, even when the sample size is as small as $n = 100$ and the dimension is as high as $d = 20$. Simulation results also suggest that the proposed tests have good power against high-frequency alternatives. One empirical application is studied in Section 7. Section 8 concludes. Some backgrounds on the RKHS is presented in Appendix A. A concise introduction of RKHS can be found in Carrasco et al. (2007), while for more comprehensive surveys on this subject, see, e.g., Hofmann et al. (2008); Paulsen and Raghupathi (2016).

2. Main Idea, Benefits and Challenges

2.1 Expressing the Conditional Moment Restrictions in RKHS

Let $Z = (Y, X^\top)$ be a random vector taking values in $\mathcal{Z} \subseteq \mathbb{R}^{1+d}$ with distribution P_Z, X a random vector taking values in $\mathcal{X} \subseteq \mathbb{R}^d$ with distribution P_X, and $\Theta \in \mathbb{R}^r$ a parameter space. Typically, Y represents the real-valued dependent (or response) variable, X is the explanatory variable. Under $\mathbb{E}|Y| < \infty$, it is well-known that the regression function $\mathbb{E}(Y|X)$ is well-defined and is the 'best' prediction of Y given X, in mean square sense. In empirical studies, it is common to consider the following expression:

$$\begin{aligned} Y &= \mathbb{E}_{\theta_0}(Y|X) + \varepsilon \\ &= \mathcal{M}(X;\theta_0) + \varepsilon \end{aligned}$$

We are interested in testing the moment restriction models where the only information about the unknown parameter $\theta_0 \in \Theta$ is a set of conditional moment restrictions:

$$\mathcal{E}(X;\theta_0) = \mathbb{E}(\varepsilon(Z;\theta_0)|X) = \mathbb{E}(Y - \mathcal{M}(X;\theta_0) \mid X) = 0 \quad P_X\text{-}a.s., \tag{1}$$

here, $\varepsilon : \mathcal{Z} \times \Theta \to \mathbb{R}$ is the generalized residual functions whose functional form are known up to the parameter $\theta \in \Theta$.

Given an i.i.d sample $\{x_i, y_i; i = 1, \ldots, n\}$ drawn from a distribution P_Z, our goal is to conduct specification testing:

$$\begin{aligned} &H_0 : \mathcal{E}(X; \theta_0) = 0 \quad P_X\text{-}a.s. \\ &H_1 : \mathcal{E}(X; \theta_0) \neq 0 \quad P_X\text{-}a.s. \quad \forall \theta \in \Theta \end{aligned} \tag{2}$$

where θ_0 has a consistent estimator $\hat{\theta}$. To do so, we follow the integrated conditional moment (ICM) approach, which converts the constraint on the conditional expectation to infinite and parametric unconditional orthogonality restrictions. Let $\mathcal{H}$ be a set of measurable functions on $\mathcal{X}$, then

$$\mathcal{E}(X; \theta_0) = 0 \Leftrightarrow \mathbb{E}\left(\varepsilon(Z; \theta_0) h(X, t)\right) = 0, P_X\text{-}a.s. \quad \forall t \in \mathcal{T}, \quad h \in \mathcal{H} \tag{3}$$

where $\mathcal{T}$ is some proper space. For sufficient conditions on the family $\mathcal{H}$ to satisfy (3), see Bierens and Ploberger (1997); Escanciano (2006b). In the context of this work, $\mathcal{H}$ must consist of infinitely many instruments for the conditional moment test to be consistent against all alternatives.

Equivalently, any $\theta_0 \in \Theta$ that satisfies (3) must also satisfy the *maximum moment restriction* (MMR) (Muandet et al., 2020):

$$\sup_{h \in \mathcal{H}} \|\mathbb{E}\left(\varepsilon(Z; \theta_0) h(X, t)\right)\|_2^2 = 0 \tag{4}$$

However, the sup operator makes it hard to optimize (4). We resolve this issue by restricting $\mathcal{H}$ to be a unit ball in a RKHS. To express (4) using the RKHS, let $h : \mathcal{X} \to \mathbb{R}$, and $\mathcal{H}(k)$ be the RHKS of functions on $\mathcal{X}$ with reproducing kernel k. The subsequent analyses rely on the following assumptions:

- (A1) The random vector (X, Z) is a strictly stationary process with probability measure P_{XZ}.

- (A2) Some regularity conditions. (i) the function $\varepsilon : \mathcal{Z} \times \Theta \to \mathbb{R}$ is continuous on Θ for each $z \in \mathcal{Z}$; (ii) $\mathcal{E}(x; \theta)$ exists and is finite for every $\theta \in \Theta$ and $x \in \mathcal{X}$ for which $P_X(x) > 0$; (iii) $\mathcal{E}(x; \theta)$ is continuous on Θ for all $x \in \mathcal{X}$ for which $P_X(x) > 0$.

- (A3) There is a unique $\theta_0 \in \Theta^\circ$ for which $\mathcal{E}(X; \theta_0) = 0, a.s.$, and $P(\mathcal{E}(X; \Theta) = 0) < 1$ for all $\theta \neq \theta_0$, where Θ° is the interior of Θ.

- (A4) The kernel $k(\cdot, \cdot)$ is *integrally strictly positive definite* (ISPD), continuous and bounded, i.e., $\sup_{x \in \mathcal{X}} \sqrt{k(x, x)} < \infty$.

Assumptions A1 and A2 are regular conditions appeared in most literature. Assumption A3 is a global identification assumption, and Assumption A4 put restrictions on the kernel k and is essential for the identification of the model. An ISPD kernel satisfies

$$\int\int_{\mathcal{X}} f(x)k(x,x')f(x')dxdx' > 0, \quad \forall \|f\|_2 \neq 0$$

where $\|\cdot\|_2$ denotes the L_2 norm.

Define an operator $\mathcal{C}_\theta : \mathcal{H}(k) \to \mathbb{R}$ that takes an instrument $h \in \mathcal{H}(k)$ as input and returns the corresponding moment restrictions:

$$\mathcal{C}_\theta h = \mathbb{E}_{XZ}\left(\varepsilon(Z;\theta)h(X)\right)$$

By the reproducing property of the RKHS, we have

$$h(x) = \langle h, \phi_x(\cdot)\rangle_{\mathcal{H}(k)}$$

where $\phi_x(\cdot) = k(x,\cdot)$ is the feature map of this RKHS with $k(x,x') = \langle \phi_x(\cdot), \phi_{x'}(\cdot)\rangle_{\mathcal{H}(k)}$.

By Riesz's representation theorem, one can show that

$$\mathcal{C}_\theta h = \langle h, \boldsymbol{\mu}_\theta\rangle_{\mathcal{H}(k)} \tag{5}$$

where

$$\boldsymbol{\mu}_\theta = \mathbb{E}_X\left(\mathcal{E}(x;\theta)\phi_x(\cdot)\right) = \mathbb{E}_{XZ}\left(\varepsilon(Z;\theta)\phi_X(\cdot)\right) \tag{6}$$

is called the *Conditional Moment Embedding* (CME) (Muandet et al., 2017, 2020). Its verification can be found in Appendix C. The idea of CME is to extend the feature map ϕ to the space of probability distribution P_X and the space of conditional moment restrictions $\mathcal{E}(X;\theta)$ by representing both elements as a mean function. Through Equation (6), most RKHS methods can therefore be extended to conditional moment restrictions.

Remark. Since $\phi_x(\cdot)$ takes values in the RKHS, the integral $\int \varepsilon(z;\theta)\phi_x(\cdot)dP_{XZ}(x,z)$ should be interpreted as a Bochner integral (see Dinculeanu (2000) for the definition of the Bochner integral).

This representation is useful. Given ISPD kernels, the CME (or equivalently, the MMR) captures all information about the conditional moment restrictions. In other words, $\boldsymbol{\mu}_\theta$ is injective, implying that for any $\theta_1, \theta_2 \in \Theta$, $\mathcal{E}(x;\theta_1) = \mathcal{E}(x;\theta_2)$ for P_X-almost surely if and only if $\boldsymbol{\mu}_{\theta_1} = \boldsymbol{\mu}_{\theta_2}$. An important consequence is $\|\boldsymbol{\mu}_\theta\|^2_{\mathcal{H}(k)} \geq 0$ and $\|\boldsymbol{\mu}_\theta\|^2_{\mathcal{H}(k)} = 0$ if and only if $\theta = \theta_0$. See, e.g., Muandet et al. (2020) for a detailed discussion.

To summarize so far, the MMR condition in (4) then can be written as

$$\sup_{\|h\|_{\mathcal{H}(k)}\leq 1} \|\mathbb{E}\left(\varepsilon(Z;\theta_0)h(X,t)\right)\|_2^2 = \|\mathcal{C}_\theta\|^2 = \|\boldsymbol{\mu}_{\theta_0}\|^2_{\mathcal{H}(k)}$$

and the original null hypothesis is equivalent to

$$H_0 : \|\boldsymbol{\mu}_{\theta_0}\|^2_{\mathcal{H}(k)} = 0, \quad P_x\text{-}a.s$$

To simplify notation, let $\mathbb{M}^2(\theta_0) = \|\boldsymbol{\mu}_{\theta_0}\|^2_{\mathcal{H}(k)}$, and further notice that

$$\begin{aligned}
\mathbb{M}^2(\theta_0) &= \langle \mathbb{E}_{XZ}(\varepsilon(Z;\theta)\phi_X(\cdot)), \mathbb{E}_{XZ}(\varepsilon(Z;\theta)\phi_X(\cdot))\rangle_{\mathcal{H}(k)} \\
&= \mathbb{E}_{XZ}\left(\langle \varepsilon(Z;\theta)\phi_X(\cdot), \mathbb{E}_{XZ}(\varepsilon(Z;\theta)\phi_X(\cdot))\rangle_{\mathcal{H}(k)}\right) \\
&= \mathbb{E}\left(\left\langle \varepsilon(Z;\theta)\phi_X(\cdot), \varepsilon(Z';\theta)\phi_{X'}(\cdot)\right\rangle_{\mathcal{H}(k)}\right) \\
&= \mathbb{E}\left(\varepsilon(Z;\theta_0)k(X,X')\varepsilon(Z';\theta_0)\right)
\end{aligned}$$

Then, given a consistent estimator $\hat{\theta}$, we propose a simple test statistic $n\widehat{\mathbb{M}}^2_n(\hat{\theta})$ as

$$n\widehat{\mathbb{M}}^2_n(\hat{\theta}) = \frac{n}{n(n-1)} \sum_{1\leq i\neq j\leq n} \varepsilon(z_i;\hat{\theta})k(x_i,x_j)\varepsilon(z_j;\hat{\theta})$$

The asymptotic distributions of $n\widehat{\mathbb{M}}^2_n(\hat{\theta})$ under different hypotheses are complicated due to the presence of the estimator $\hat{\theta}$. In Section 3, we propose a projection-based test that has the ability to eliminate this estimation effect. Hence, we will not study the asymptotic distributions of $n\widehat{\mathbb{M}}^2_n(\hat{\theta})$ in detail.

2.2 Benefits of using RKHS Techniques: Insensitive to Dimension and Massively Generate ICM Tests

2.2.1 Dimension Reduction

Let $\widehat{\boldsymbol{\mu}}_{\hat{\theta}} = 1/n\sum_{i=1}^n \varepsilon(z_i;\hat{\theta})\phi_{x_i}(\cdot) \in \mathcal{H}(k)$ be an estimator of $\boldsymbol{\mu}_{\theta_0}$, and suppose $\hat{\theta} - \theta_0 = O_p(1/\sqrt{n})$. Let $g(z;\theta) = \nabla_\theta \varepsilon(z;\theta)$ be the first derivative of $\varepsilon(z;\theta)$, and $\bar{\theta} = \gamma\theta_0 + (1-\gamma)\hat{\theta}, \gamma \in (0,1)$. Notice that

$$\widehat{\boldsymbol{\mu}}_{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^n \varepsilon(z_i;\theta_0)\phi_{x_i}(\cdot) + O_p(1/\sqrt{n})^\top \frac{1}{n}\sum_{i=1}^n g(z_i;\bar{\theta})\phi_{x_i}(\cdot)$$

and

$$\begin{aligned}
\|\widehat{\boldsymbol{\mu}}_{\hat{\theta}} - \boldsymbol{\mu}_{\theta_0}\|_{\mathcal{H}(k)} &= \|\widehat{\boldsymbol{\mu}}_{\theta_0} + O_p(1/\sqrt{n})^\top \frac{1}{n}\sum_{i=1}^n g(z_i;\bar{\theta})\phi_{x_i}(\cdot) - \boldsymbol{\mu}_{\theta_0}\|_{\mathcal{H}(k)} \\
&\leq \|\widehat{\boldsymbol{\mu}}_{\theta_0} - \boldsymbol{\mu}_{\theta_0}\|_{\mathcal{H}(k)} + O_p(1/\sqrt{n})\sqrt{\frac{1}{n^2}\sum_{i,j=1}^n g^\top(z_i;\bar{\theta})k(x_i,x_j)g(z_j;\bar{\theta})} \\
&= \|\widehat{\boldsymbol{\mu}}_{\theta_0} - \boldsymbol{\mu}_{\theta_0}\|_{\mathcal{H}(k)} + O_p(1/\sqrt{n})
\end{aligned}$$

Furthermore, Muandet et al. (2017); Tolstikhin et al. (2017) show that

$$\|\widehat{\boldsymbol{\mu}}_{\theta_0} - \boldsymbol{\mu}_{\theta_0}\|_{\mathcal{H}(k)} = O_p(1/\sqrt{n})$$

This observation states that $\widehat{\mu}_{\hat{\theta}}$ converges in the RKHS norm in a way that is independent of the dimension of (X, Z). This is an appealing property since estimation and inference based on $\widehat{\mu}_{\hat{\theta}}$ is less sensitive to the curse of dimensionality.

The following isomorphic result helps us to investigate where the dimensionality is 'hiding'. Let $L_2(\mathbb{R}^{1+d},\Pi)=\{f:\mathbb{R}^{1+d}\to\mathbb{R}, s.t.\|f\|=\left(\int|f|^2d\Pi\right)^{1/2}<\infty\}$.

Since bothe the RKHS and the $L_2(\mathbb{R}^{1+d},\Pi)$ are separable and infinite dimensional Hilbert spaces, these two spaces are isometrically isomorphic, i.e., there exists a one-to-one linear mapping $J:\mathcal{H}(k)\to L_2(\mathbb{R}^{1+d},\Pi)$ such that

$$\langle J(f),J(g)\rangle_{L_2(\mathbb{R}^{1+d},\Pi)}=\langle f,g\rangle_{\mathcal{H}(k)},\quad f,g\in\mathcal{H}(k)$$

See Carrasco et al. (2007) for details.

Notice that the V-statistic version of $n\widehat{\mathbb{M}}_n^2(\hat{\theta})$ can be thought as

$$\begin{aligned}\frac{n}{n^2}\sum_{i,j=1}^{n}\varepsilon(z_i;\hat{\theta})k(x_i,x_j)\varepsilon(z_j;\hat{\theta})&=\left\langle\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon(z_i;\hat{\theta})\phi_{x_i}(\cdot),\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon(z_j;\hat{\theta})\phi_{x_j}(\cdot)\right\rangle_{\mathcal{H}(k)}\\&=\left\langle\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon(z_i;\hat{\theta})J(\phi_{x_i}(\cdot)),\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon(z_j;\hat{\theta})J(\phi_{x_j}(\cdot))\right\rangle_{L_2(\mathbb{R}^{1+d},\Pi)}\end{aligned}\tag{7}$$

where the first equality is a consequence of the reproducing property, and the last equality arises from the isometrically isomorphic relationship. Thus, the proposed test statistic $n\widehat{\mathbb{M}}_n^2(\hat{\theta})$ is an U-statistic version of an ICM test with a weighting function $J(\phi_x(\cdot))$.

Furthermore, when the kernel is chosen to be 'shift-invariant', i.e., the kernel is solely depends on the difference of its arguments,

$$k(x,x')=\psi(x-x')$$

a more specific ICM structure is revealed by the following characterization, which is due to Bochner (1933), see also Rudin (2017). We state it in the form given by Wendland (2004).

Theorem 1 *(Bochner). Let $k(x,x')=\psi(x-x')$ be a shift-invariant kernel for continuous function $\psi:\mathbb{R}^d\to\mathbb{C}$. Then ψ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure Λ on $\mathbb{R}^d$:*

$$\psi(t)=\int_{\mathbb{R}^d}\exp(-i\langle t,\omega\rangle)d\Lambda(\omega)$$

for $t\in\mathbb{R}^d$.

One may normalize ψ such that $\psi(0)=1$, in which case Λ is a probability measure and ψ is its characteristic function. For example, if Λ is a normal distribution of the form $(2\pi/\sigma^2)^{-d/2}e^{-\frac{\sigma^2\|\omega\|^2}{2}}d\omega$, then the corresponding ISPD kernel is the Gaussian $\exp(-\|t\|^2/2\sigma^2)$. By applying Bochner's theorem, one can show that, see, e.g., Fan and Li (2000); Muandet et al. (2020)

$$\mathbb{M}^2(\theta)=(2\pi)^{-d/2}\int_{\mathbb{R}^d}\mathbb{E}\left(\varepsilon(Z;\theta)\exp(-i\omega^\top X)\right)^2d\Lambda(\omega)$$

where Λ is a Fourier transform of the kernel k.

Several commonly studied ICM tests are indeed in a form of $n\widehat{\mathbb{M}}_n^2(\hat{\theta})$ (V-statistic version). For example, the ICM test of Birens with exponential weighting function $\exp(i\omega^\top x)$ and a uniform integral measure can be stated as

$$ICM_n = \frac{n}{n^2}\sum_{j,k=1}^{n} \varepsilon(z_j;\hat{\theta})\varepsilon(z_k;\hat{\theta})\exp\left(-\frac{1}{2}\|x_j - x_k\|^2\right)$$

where the kernel is chosen as the Gaussian RBF $k(x,x') = \exp\left(-\|x-x'\|_2^2/2\sigma^2\right), \sigma = 1$.

Another popular ICM test is Escanciano's $PCvM_n$ (Escanciano, 2006a) with a weighting function of $\mathbb{I}\{\omega^\top x \leq u\}$ and an empirical distribution as the integral measure, its analytic closed form is:

$$PCvM_n = \frac{n}{n^2}\sum_{j\neq k}^{n} \varepsilon(z_j;\hat{\theta})\varepsilon(z_k;\hat{\theta})\left(\frac{1}{n}\sum_{r=1}^{n} B_{jkq}^{(0)}\frac{\pi^{(d/2)-1}}{\Gamma(\frac{d}{2}+1)}\right)$$

where $\Gamma(\cdot)$ denotes the gamma function, $B_{jkq}^{(0)}$ is the complementary angle between $(x_j - x_q)$ and $x_k - x_q$, and is defined as

$$B_{jkq}^{(0)} = \left|\pi - arcos\left(\frac{(x_j - x_q)^\top(x_k - x_q)}{|x_j - x_q||x_k - x_q|}\right)\right|$$

Equation (7) and Bochner's theorem state that both weighting functions and integral measures are implicitly determined by the kernel.

2.2.2 Generate ICM Tests, Massively and Cheaply

Constructing ICM tests from weighting functions and integral measures is difficult, as one needs to perform the integration. Observe that each kernel corresponds to an ICM test, and the RKHS representation provides a cheap way to massively produce ICM tests by constructing kernels from existing kernels. To begin with, we first introduce some commonly used kernels listed in Table 1. For more examples, see, e.g., Muandet et al. (2017); Steinwart (2001); Steinwart and Christmann (2008).

Table 1: Various characterizations of known kernels

Kernel Function	$k(x,x')$	Domain $\mathcal{X}$	Characteristic	Shift Invariant	ISPD
Gaussian	$\exp\left(-\gamma\|x-x'\|_2^2\right), \gamma > 0$	$\mathbb{R}^d$	✓	✓	✓
Laplacian	$\exp\left(-\|x-x'\|_1/\sigma\right), \sigma > 0$	$\mathbb{R}^d$	✓	✓	✓
Inverse Multiquadric	$(c^2 + \|x-x'\|_2^2)^{-\gamma}, c, \gamma > 0$	$\mathbb{R}^d$	✓	✓	✓
Exponential	$\exp(\sigma\langle x, x'\rangle), \sigma > 0$	Compact sets of $\mathbb{R}^d$	✓	✗	✓
Matern	$2^{1-\nu}\Gamma^{-1}(\nu)\left(\sqrt{2\nu}/\rho\|x-x'\|_2\right)^\nu \kappa_\nu\left(\sqrt{2\nu}/\rho\|x-x'\|_2\right)$	$\mathbb{R}^d$	✓	✓	✓
Infinite Polynomial	$(1-\langle x,x'\rangle)^{-\alpha}, \alpha > 0$	$\{x \in \mathbb{R}^d : \|x\|_2 < 1\}$	✓	✗	✓

Notes: In the Matern kernel, $\Gamma(\cdot)$ is the Gamma function, κ_ν is the modified Bessel function of the second type, ν, ρ are non-negative parameters. When $\nu \to \infty$, it becomes equivalent to the Gaussian kernel, when $\nu = 1/2$, it reduces to the Laplacian kernel.

The following lemmas describe ways of constructing new ISPD kernels.

Lemma 2 *Let $a \geq 0$, and k, k_1 and k_2 be ISPD kernels on X. Then ak and $k_1 + k_2$ are also ISPD kernels on X.*

This lemma states that the set of ISPD kernels is a convex cone, its proof is trivial and will not be discussed here.

Lemma 3 *A shift variant ISPD kernel, $\tilde{k}$ can be obtained from a shift invariant ISPD kernel, k, as*

$$\tilde{k}(x, x') = f(x)k(x, x')f(x')$$

where $f : \mathcal{X} \longrightarrow \mathbb{R}$ is a bounded continuous function.

Proof See Sriperumbudur et al. (2010). ∎

This lemma states that one can generate new ISPD kernels through a conformal mapping, i.e., a transformation that preserves angles locally.

The next lemma is based on the fact that all bounded continuous shift-invariant kernels, if they are characteristic, are also ISPD. A measurable and bounded kernel, k is said to be characteristic if

$$\mathbb{P} \longrightarrow \int_{\mathcal{X}} k(\cdot, x) d\mathbb{P}(x)$$

is injective, that is, $\mathbb{P}$ is embedded to a unique element in $\mathcal{H}(k)$. The above mentioned shift-invariant kernels (i.e., the Gaussian, the Laplacian, the IMQ and the Matern) are all characteristic kernels.

Lemma 4 *Let k, k_1 and k_2 be a bounded continuous shift-invariant kernel on $\mathbb{R}^d$. Suppose k is characteristic and $k_2 \neq 0$, then $k + k_1$ and $k \times k_2$ are characteristic.*

Proof See Sriperumbudur et al. (2010). ∎

Escanciano (2009) has shown that ICM tests only have substantial local power against alternatives in a finite-dimensional space, and there is only one direction with the highest asymptotic local power. This best direction depends on the weighting function, the integrated measure, the true model and DGP. Since a kernel embodies the weighting function and the integral measure, it also affects the directions in which the corresponding ICM test has substantial power. Hence, different kernels would have different power properties, and it is desirable to have as many ICM tests as possible.

Choosing kernel is important. Some kernels will gradually reduce to a constant function as the dimension d of X increases, making corresponding tests distorting size as well as losing power in almost all directions. Considering the Birens' test, which corresponds to the Gaussian kernel with Euclidean distance and parameter $\gamma = 1/2$. When d is large, the corresponding kernel matrix K_{ij} (also known as the Gram matrix, see Appendix A for details) becomes close to the identity matrix[1]. Nevertheless, one could 'slow down' the decay rate by changing the parameter inside the Gaussian kernel, we will demonstrate this point in the simulation exercises.

2.3 Challenges when Using the Simple Statistic

We call the test statistic

$$n\widehat{\mathbb{M}}_n^2(\hat{\theta}) = \frac{n}{n(n-1)} \sum_{1\leq i\neq j\leq n} \varepsilon(z_i;\hat{\theta})k(x_i,x_j)\varepsilon(z_j;\hat{\theta})$$

the simple statistic because it simply replaces unknown parts by their empirical counterparts. There are several potential drawbacks of using this simple statistic. First, if the estimator $\hat{\theta}$ is standard in a sense that $O_p(\|\hat{\theta}-\theta_0\|) = O_p(1/\sqrt{n})$, then the distributions of the test statistic would depend on how $\hat{\theta}$ is estimated. Furthermore, in most literature, to establish the limiting distribution of $n\widehat{\mathbb{M}}_n^2(\hat{\theta})$ under the null, an asymptotic linear representation for $\sqrt{n}(\hat{\theta}-\theta_0)$ is often required, see, e.g., Delgado et al. (2006); Escanciano (2006a).

Second, since the limiting distribution of $n\widehat{\mathbb{M}}_n^2(\hat{\theta})$ is non-pivotal, a bootstrap procedure is needed to calculate critical values. However, the presence of $\hat{\theta}$ often requires a case-by-case complicated parametric bootstrap procedure.

Finally, and perhaps more interestingly, certain 'non-standard' estimators $\hat{\theta}$ with slower than $1/\sqrt{n}$ rate of convergence are ruled out, as in these cases, $\lim_{n\to\infty} O_p(\sqrt{n}\|\hat{\theta}-\theta_0\|) \to \infty$.

To deal with this problem, one could try to find suitable transformations on the kernel to eliminate the 'estimation effect'. Most literature are based on the empirical process and are adopting two different transformation approaches. The first approach consists of martingale transformation of the empirical process, see, for instance, Delgado and Stute (2008); Khmaladze (1982, 1993); Koul and Stute (1999). However, the martingale transformation can be quite complicated even for some conventional econometrics models. More importantly, this transformation is based on a sequence of iterative regressions, the inversion of the projection matrix could be unstable, which will ultimately affect the sampling performance.

The second approach is based on the idea of projecting the weighting function $h(X,t)$ onto a tangent space of nuisance parameters, see, for example, Bickel et al. (2006); Escanciano and Goh (2014); Neyman (1959); Sant'Anna and Song (2020); Sant'Anna and Song (2019). This approach is relatively easier to implement and requires weaker conditions than the Khmaladze transformation. In this study, we extend this projection idea to kernels.

1. Identity matrix for the V-statistic version of the test, and the kernel matrix is close to a zero matrix, where all elements are zeros, if one uses the U-statistic version.

3. A Projection Test Statistic and its Asymptotic Results

3.1 The Projection Test Statistic

Let $g(z;\theta), \bar{\theta}$ are defined as before, the simple statistic can be expanded as

$$\begin{aligned} n\widehat{\mathbb{M}}_n^2(\hat{\theta}) &= \frac{n}{n(n-1)} \sum_{1\le i\ne j\le n} \Big(\varepsilon(z_i;\theta_0) + g^\top(z_i;\bar{\theta})(\hat{\theta}-\theta_0)\Big)\, k(x_i,x_j)\, \Big(\varepsilon(z_j;\theta_0) + g^\top(z_j;\bar{\theta})(\hat{\theta}-\theta_0)\Big) \\ &= nA_{1,n}(k) + 2\sqrt{n}A_{2,n}(k)\sqrt{n}(\hat{\theta}-\theta_0) + \sqrt{n}(\hat{\theta}-\theta_0)^\top A_{3,n}(k)\sqrt{n}(\hat{\theta}-\theta_0) \end{aligned}$$

where

$$\begin{aligned} A_{1,n}(k) &= \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \varepsilon(z_i;\theta_0)k(x_i,x_j)\varepsilon(z_j;\theta_0) \\ A_{2,n}(k) &= \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \varepsilon(z_i;\theta_0)k(x_i,x_j)g^\top(z_j;\bar{\theta}) \\ A_{3,n}(k) &= \left(\frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} g(z_i;\bar{\theta})k(x_i,x_j)g^\top(z_j;\bar{\theta})\right) \end{aligned}$$

In order to eliminate the estimation effect $\sqrt{n}(\hat{\theta}-\theta_0)$, we need to find a $k_p(\cdot,\cdot)$ such that

$$\mathbb{E}A_{2,n}(k_p) = \mathbb{E}\left(\varepsilon(Z;\theta_0)g(Z';\theta_0)k_p(X,X'\right) = \mathbf{0}$$

and

$$\mathbb{E}A_{3,n}(k_p) = \mathbb{E}\Big(g(Z;\theta_0)k_p(X,X')g(Z';\theta_0)^\top\Big) = 0$$

One possibility is

$$\begin{aligned} k_p(x,x') = k(x,x') &- g^\top(z;\theta_0)\Gamma_{\theta_0}^{-1}\mathbb{E}_{(X,Z)}\left(g(Z;\theta_0)k(X,x')\right) \\ &- g^\top(z';\theta_0)\Gamma_{\theta_0}^{-1}\mathbb{E}_{(X',Z')}\left(g(Z';\theta_0)k(X',x)\right) \\ &+ g^\top(z;\theta_0)\Gamma_{\theta_0}^{-1}\mathbb{E}\Big(g(Z;\theta_0)k(X,X')g^\top(Z';\theta_0)\Big)\Gamma_{\theta_0}^{-1}g(z';\theta_0) \end{aligned} \tag{8}$$

where $\Gamma_\theta = \mathbb{E}\left(g(Z;\theta)g^\top(Z;\theta)\right)$, and

$$\mathbb{E}_{(X,Z)}\left(g(Z;\theta_0)k(X,x')\right) = \mathbb{E}\left(g(Z;\theta_0)k(X,x')|X'=x'\right)$$

The corresponding test statistic $n\widehat{\mathbb{M}}_p^2(\hat{\theta})$ is specified as:

$$n\widehat{\mathbb{M}}_p^2(\hat{\theta}) = \frac{n}{n(n-1)} \sum_{1\le i\ne j\le n} \varepsilon(z_i;\hat{\theta})\hat{k}_p(x_i,x_j)\varepsilon(z_j;\hat{\theta}) \tag{9}$$

where $\hat{k}_p(\cdot,\cdot)$ is the empirical counterpart of $k_p(\cdot,\cdot)$.

In the subsequent contents, we will explain (1) Where does $k_p(\cdot,\cdot)$ come from? (2) What are the properties of $k_p(\cdot,\cdot)$, and (3) Does

$$\mathcal{E}(X;\theta_0) = 0 \Leftrightarrow \mathbb{E}\left(\varepsilon(Z;\theta_0)k_p(X,X')\varepsilon(Z';\theta_0)\right) = 0, \quad P_x\text{-}a.s.$$

hold?

A Projection Interpretation.

In the conventional ICM framework, the canonical way to 'swipe out' the estimation effect is to project the weighting function onto the tangent space of nuisance parameters, see, for instance, Escanciano and Goh (2014); Sant'Anna and Song (2020); Sant'Anna and Song (2019). We adopt the same approach to the feature map $\phi_x(\cdot)$ of k. Define a projection operator $\mathcal{P}$ that takes a value in $\mathcal{H}(k)$ and delivers the projected feature map $\mathcal{P}\phi_x(\cdot)$:

$$\mathcal{P}\phi_x(\cdot) = \phi_x(\cdot) - g^\top(z;\theta_0)\Gamma_{\theta_0}^{-1}\mathbb{E}_{XZ}\left(g(Z;\theta_0)\phi_X(\cdot)\right) \tag{10}$$

The intuition behind (10) is simple. First, note that $\Gamma_{\theta_0}^{-1}\mathbb{E}_{XZ}\left(g(Z;\theta_0)\phi_X(\cdot)\right)$ is the vector of linear projection coefficients of regressing $\phi_x(\cdot)$ on the score function $g(z;\theta_0)$. Thus, $g^\top(z;\theta_0)\Gamma_{\theta_0}^{-1}\mathbb{E}_{XZ}\left(g(Z;\theta_0)\phi_X(\cdot)\right)$ is the best linear predictor of $\phi_x(\cdot)$ given $g(z;\theta_0)$, and equation (10) is nothing more than the associated projection error, which, by definition, is orthogonal to $g(z;\theta_0)$.

Observe that $\mathcal{P}$ is a linear operator, it follows the construction of RKHS that $\mathcal{P}\phi_x(\cdot) \in \mathcal{H}(k)$. This can be verified by checking its reproducing property (see Appendix C):

$$\begin{aligned}\langle \mathcal{P}\phi_x(\cdot), \phi_{x'}(\cdot)\rangle_{\mathcal{H}(k)} &= \mathcal{P}\phi_x(x') \\ &= k(x,x') - g^\top(z;\theta_0)\Gamma_{\theta_0}^{-1}\mathbb{E}_{XZ}\left(g(Z;\theta_0)k(X,x')\right)\end{aligned} \tag{11}$$

k_p is then constructed from

$$k_p(x,x') = \langle \mathcal{P}\phi_x(\cdot), \mathcal{P}\phi_{x'}(\cdot)\rangle_{\mathcal{H}(k)} \tag{12}$$

It can be understood as the 'residual square' of $\phi_x(\cdot)$.

Properties of K_p.

Note that the operator $\mathcal{P}$ is an orthogonal projection operator in the Hilbert space of $L_2(\mathbb{R}^{1+d}, P_{(X,Z)})$ but not necessarily in the RKHS $\mathcal{H}(k)$. Thus, in general

$$\langle \phi_x(), \mathcal{P}\phi_{x'}(\cdot)\rangle_{\mathcal{H}(k)} \neq \langle \mathcal{P}\phi_x(), \mathcal{P}\phi_{x'}(\cdot)\rangle_{\mathcal{H}(k)}$$

and

$$\langle \phi_x(), \mathcal{P}\phi_{x'}(\cdot)\rangle_{\mathcal{H}(k)} \neq \langle \mathcal{P}\phi_x(), \phi_{x'}(\cdot)\rangle_{\mathcal{H}(k)}$$

Nevertheless, $\mathcal{P}$ is idempotent,

$$\mathcal{P}\mathcal{P}\phi_x(\cdot) = \mathcal{P}\phi_x(\cdot) - \mathcal{P}g^\top(z;\theta_0)\Gamma_{\theta_0}^{-1}\mathbb{E}_{XZ}\left(g(Z;\theta_0)\phi_X(\cdot)\right) = \mathcal{P}\phi_x(\cdot)$$

To investigate the positive definiteness of $k_p(\cdot,\cdot)$, it is tantamount to verifying the sign of

$$\int_{\mathcal{X}\times\mathcal{X}} f(x)k_p(x,x')f(x')dxdx', \quad f \in L_2$$

but the above equation can be written as

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{X}} \left\langle f(x)\mathcal{P}\phi_x(\cdot), f(x')\mathcal{P}\phi_{x'}(\cdot)\right\rangle_{\mathcal{H}(k)} dxdx' \\
&= \left\langle \int_{\mathcal{X}} f(x)\mathcal{P}\phi_x(\cdot)dx, \int_{\mathcal{X}} f(x')\mathcal{P}\phi_{x'}(\cdot)dx' \right\rangle_{\mathcal{H}(k)} \\
&= \left\| \int_{\mathcal{X}} f(x)\mathcal{P}\phi_x(\cdot)dx \right\|^2_{\mathcal{H}(k)} \geq 0
\end{aligned}$$

where the last equality comes from the independence between x and x'. The Moore-Aronszajn Theorem states that this positive definite kernel $k_p(\cdot,\cdot)$ is associated with an unique RKHS $\mathcal{H}(k_p)$.

Let $\mathcal{P}^1$ with $\mathcal{P}^1\phi_x(\cdot) = \phi_x(\cdot) - \mathcal{P}\phi_x(\cdot)$ be another orthogonal projection operator, and by properties of $\mathcal{P}^1$, we have

$$\begin{aligned}
\|\mathcal{P}^1\phi_x(\cdot)\|^2_{L_2(\mathbb{R}^{1+d},P_{(X,Z)})} &= \left\langle \mathcal{P}^1\phi_x(\cdot), \mathcal{P}^1\phi_x(\cdot)\right\rangle_{L_2(\mathbb{R}^{1+d},P_{(X,Z)})} \\
&= \left\langle \mathcal{P}^1\phi_x(\cdot), \phi_x(\cdot)\right\rangle_{L_2(\mathbb{R}^{1+d},P_{(X,Z)})} \\
&\leq \|\mathcal{P}^1\phi_x(\cdot)\|_{L_2(\mathbb{R}^{1+d},P_{(X,Z)})} \|\phi_x(\cdot)\|_{L_2(\mathbb{R}^{1+d},P_{(X,Z)})}
\end{aligned}$$

Thus,

$$\|\mathcal{P}^1\phi_x(\cdot)\|_{L_2(\mathbb{R}^{1+d},P_{(X,Z)})} \leq \|\phi_x(\cdot)\|_{L_2(\mathbb{R}^{1+d},P_{(X,Z)})}$$

By the isometrically isomorphic relationship between $\mathcal{H}(k)$ and $L_2(\mathbb{R}^{1+d}, P_{(X,Z)})$, we further have

$$\|\mathcal{P}^1\phi_x(\cdot)\|_{\mathcal{H}(k)} \leq \|\phi_x(\cdot)\|_{\mathcal{H}(k)} \tag{13}$$

with equality holds if $\phi_x(\cdot) \in span\{g(z;\theta_0) : z \in \mathcal{Z}\}$.

Thus, as long as $\phi_x(\cdot) \notin span\{g(z;\theta_0) : z \in \mathcal{Z}\}$,

$$\|\mathcal{P}\phi_x(\cdot)\|_{\mathcal{H}(k)} = \|\phi_x(\cdot) - \mathcal{P}^1\phi_x(\cdot)\|_{\mathcal{H}(k)} > 0, \quad \forall x \in \mathcal{X}$$

$$\left\| \int_{\mathcal{X}} f(x)\mathcal{P}\phi_x(\cdot)dx \right\|^2_{\mathcal{H}(k)} > \left\| \int_{\mathcal{X}} f(x)dx \inf_{x\in\mathcal{X}} \mathcal{P}\phi_x(\cdot) \right\|^2_{\mathcal{H}(k)} \geq 0$$

and $k_p(\cdot,\cdot)$ is an ISPD kernel.

(Almost) Equivalence between $\mathcal{E}(X;\theta_0)$ and $\mathbb{E}\left(\varepsilon(Z;\theta_0)k_p(X,X')\varepsilon(Z';\theta_0)\right)$

Similar to the story presented in Section 2, we show $\mathcal{E}(X;\theta)$ is almost injective to a conditional moment embedding $\mu_\theta^{(p)} \in \mathcal{H}(k_p)$ and $\mathbb{E}\left(\varepsilon(Z;\theta_0)k_p(X,X')\varepsilon(Z';\theta_0)\right) = \|\mu_{\theta_0}^{(p)}\|^2_{\mathcal{H}(k_p)}$.

Redefine the operator $\mathcal{C}_\theta$ as $\mathcal{C}_\theta^{(p)} : \mathcal{H}(k_p) \to \mathbb{R}$:

$$\mathcal{C}_\theta^{(p)} h = \mathbb{E}_{XZ}\left(\varepsilon(Z;\theta)h(X)\right), \quad h \in \mathcal{H}(k_p)$$

Let $\phi_x^{(p)}(\cdot)$ be the feature map associated with $k_p(\cdot,\cdot)$. By the reproducing property, we have $h(x) = \langle h, \phi_x^{(p)}(\cdot)\rangle_{\mathcal{H}(k_p)}$ and

$$\mathcal{C}_\theta^{(p)} h = \left\langle h, \mathbb{E}_{XZ}\left(\varepsilon(Z;\theta)\phi_x^{(p)}(\cdot)\right)\right\rangle_{\mathcal{H}(k_p)} = \left\langle h, \mu_\theta^{(p)}\right\rangle_{\mathcal{H}(k_p)}$$

By Riesz's representor theorem,

$$|\mathcal{C}_\theta^{(p)}| = \|\mu_\theta^{(p)}\|_{\mathcal{H}(k_p)}$$

The following theorem states the almost injectivity between $\mathcal{E}(X;\theta)$ and $\mu_\theta^{(p)}$:

Theorem 5 *For any $\theta_1, \theta_2 \in \Theta$, assume $\mathcal{E}(x;\theta)$ is not collinear with $g(Z;\theta_0)$, then we have $\mathcal{E}(X;\theta_1) = \mathcal{E}(X;\theta_2)$ if and only if $\mu_{\theta_1}^{(p)} = \mu_{\theta_2}^{(p)}$. Consequently,*

$$\mathcal{E}(X;\theta_0) = 0 \Leftrightarrow \|\mu_{\theta_0}^{(p)}\|^2_{\mathcal{H}(k_p)} = 0 \quad P_x\text{-}a.s.$$

Proof See Appendix C. ■

Finally, by the construction of $\mu_\theta^{(p)}$, it is easy to check that

$$\|\mu_{\theta_0}^{(p)}\|^2_{\mathcal{H}(k_p)} = \mathbb{E}\left(\varepsilon(Z;\theta_0)k_p(X,X')\varepsilon(Z';\theta_0)\right)$$

3.2 Asymptotic Null Distribution

One can estimate $\mathbb{M}_p^2(\theta_0) = \|\mu_{\theta_0}^{(p)}\|^2_{\mathcal{H}(k_p)}$ by

$$\widehat{\mathbb{M}}_p^2(\hat\theta) = \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} \varepsilon(z_i;\hat\theta)\hat k_p(x_i,x_j)\varepsilon(z_j;\hat\theta)$$

Consequently, define the test statistic as $n\widehat{\mathbb{M}}_p^2(\hat\theta)$. In this subsection, we study the asymptotic properties of $n\widehat{\mathbb{M}}_p^2(\hat\theta)$ under the null hypothesis.

To derive our theoretical results, we impose additional assumptions:

- (A5). (i) $\mathbb{E}\|\varepsilon(Z;\theta_0)\|^2 < \infty$. Let $\nabla_\theta g(z;\theta)$ exist almost surely in an open neighborhood $\mathcal{N}(\theta_0)$ of θ_0. (ii) $\mathbb{E}\|g(Z;\theta_0)\| < \infty$ and $\sup_{\theta\in\mathcal{N}(\theta_0)}\|\nabla_\theta g(\cdot,\theta)\| < S(\cdot)$, with $\mathbb{E}S(Z) < \infty$, where $\|\cdot\|$ denotes either a vector or matrix norm. (iii) $\nabla_\theta g(z,\theta)$ are continuous in θ for $\theta \in \mathcal{N}(\theta_0)$ and uniformly in Z almost everywhere.

- (A6). (i) $\|\hat{\theta}-\theta_0\| = o_p(n^{-1/4})$; (ii) Γ_θ is nonsingular uniformly in $\theta \in \Theta^\circ$.

Remark. Assumption A5 contains regularity conditions for $\varepsilon(z;\theta)$, and these conditions are similar in, e.g., Delgado et al. (2006); Newey (1985); Robinson (1991). A sufficient condition for Assumption A6 (i) is $n^\varsigma\|\hat{\theta}-\theta_0\| = O_p(1)$ for some $\varsigma > 1/4$. In addition, we do not require $\hat{\theta}$ to have an asymptotically linear representation.

Lemma 6 *Under the null, we have*

$$\hat{k}_p(\cdot,\cdot) = k_p(\cdot,\cdot) + O_p(1/\sqrt{n}) + O_p(\|\hat{\theta}-\theta_0\|) \tag{14}$$

while expanding $n\widehat{\mathbb{M}}_p^2(\hat{\theta})$ around θ_0 yields

$$\begin{aligned} n\widehat{\mathbb{M}}_p^2(\hat{\theta}) =& \frac{n}{n(n-1)}\sum_{i\neq j}\varepsilon(z_i;\theta_0)k_p(x_i,x_j)\varepsilon(z_j,\theta_0) + O_p(\|\hat{\theta}-\theta_0\|) + O_p(\|\hat{\theta}-\theta_0\|^2) \\ &+ O_p(1/\sqrt{n}) \end{aligned} \tag{15}$$

Theorem 7 *Assume that $\mathbb{M}_p^2(\theta) < \infty$ for all $\theta \in \Theta$ and Assumption A6 (i) hold, under the null, we have*

$$n\widehat{\mathbb{M}}_p^2(\hat{\theta}) \xrightarrow{d} \sum_{k=1}^{\infty}\tau_k^{(p)}(W_k^2-1) \tag{16}$$

where $W_k \sim N(0,1)$, $\{\tau_k^{(p)}\}$ are eigenvalues of the operator A defined as $(A\psi)(v) = \int f(v,v')\psi(v')dP_v(v')$ for non-zero ψ, $v = (x,y)$, and $f(v,v') = \varepsilon(z;\theta_0)k_p(x,x')\varepsilon(z';\theta_0)$

Proof See, Serfling (1980). ■

3.3 Asymptotic Power

We now study the asymptotic distribution of $n\widehat{\mathbb{M}}_p^2(\hat{\theta})$ under fixed alternative and a sequence of local alternatives converging to null at a parametric rate $n^{-1/2}$.

We first consider the fixed alternative hypothesis. Observe that[2]

$$\begin{aligned} \sqrt{n}\widehat{\mathbb{M}}_p^2(\hat{\theta}) &= \frac{\sqrt{n}}{n(n-1)}\sum_{i\neq j}\varepsilon(z_i;\theta_0)k_p(x_i,x_j)\varepsilon(z_j,\theta_0) + O_p(\|\hat{\theta}-\theta_0\|) + O_p(\|\hat{\theta}-\theta_0\|^2) + O_p(1/\sqrt{n}) \\ &= \frac{\sqrt{n}}{n(n-1)}\sum_{i\neq j}\varepsilon(z_i;\theta_0)k_p(x_i,x_j)\varepsilon(z_j,\theta_0) + O_p(\|\hat{\theta}-\theta_0\|) + O_p(1/\sqrt{n}) \end{aligned}$$

2. Since both $A_{1,n}^{(p)}$ and $A_{2,n}^{(p)}$ are non-degenerate, and $A_{1,n}^{(p)}, A_{2,n}^{(p)} = O_P(1/\sqrt{n})$. $A_{1,n}^{(p)}$ and $A_{2,n}^{(p)}$ are defined in the proof of Lemma 6.

Theorem 8 *Assume that* $\mathbb{M}_p^2(\theta) < \infty$ *for all* $\theta \in \Theta$*, and* $\hat{\theta} \xrightarrow{p} \theta_1 \in \Theta$*. Under the fixed alternative, we have*

$$\sqrt{n}\left(\widehat{\mathbb{M}}_p^2(\hat{\theta}) - \mathbb{M}_p^2(\theta_1)\right) \xrightarrow{d} N(0, \sigma^2_{\theta_1,p}) \tag{17}$$

where

$$\sigma^2_{\theta_1,p} = 4\mathrm{Var}_{(X,Z)}\left(\mathbb{E}_{(X',Z')}\left(\varepsilon(Z;\theta_1)k_p(X,X')\varepsilon(Z';\theta_1)\right)\right)$$

Proof See Serfling (1980). ■

It is readily to see that for large n and fixed critical value c_α, the test power can be approximated by

$$P_{H_1}(n\widehat{M}_p^2(\hat{\theta}) > c_\alpha) \approx \Phi\left(\frac{\sqrt{n}M_P^2(\theta_1)}{\sigma_{\theta_1,p}} - \frac{c_\alpha}{\sqrt{n}\sigma_{\theta_1,p}}\right)$$

where Φ denotes the cumulative distribution function of the standard normal distribution.

Assume that n is sufficiently large, in $\sqrt{n}M_P^2(\theta_1)/\sigma_{\theta_1,p} - c_\alpha/\sqrt{n}\sigma_{\theta_1,p}$, we observe that the second term $c_\alpha/\sqrt{n}\sigma_{\theta_1,p} = O(n^{-1/2})$ going to 0 as $n \to \infty$, while the first term $\sqrt{n}M_P^2(\theta_1)/\sigma_{\theta_1,p} = O(n^{1/2})$, dominating the second. Thus, the best kernel that maximize the test power is given by

$$k^* = \arg\sup_{k\in\mathcal{K}} \frac{\sqrt{n}M_P^2(\theta_1)}{\sigma_{\theta_1,p}}$$

where $\mathcal{K}$ is a proper kernel space, e.g., $\mathcal{K} = \{\exp(-\gamma\|x - x'\|), \gamma > 0\}$.

A heuristic way to estimate k^* is to divide the sample $\{(x_i, z_i), i = 1, \ldots, n\}$ into two disjoint training and test sets, and use the training set to compute $(\widehat{M}_p(\hat{\theta})/\hat{\sigma}_{\theta_1,p})(k)$, which can be maximized by choosing the kernel parameter (e.g., in Gaussian kernel, the kernel parameter is γ). We denote the kernel that maximize $(\widehat{M}_p(\hat{\theta})/\hat{\sigma}_{\theta_1,p})(k)$ as $\hat{k}^*$. We then, use $\hat{k}^*$ and perform testing in the test set.

A similar idea has been discussed in the machine learning literature, where the test of interest is the equality of two samples, see, e.g., Gretton et al. (2012). Nevertheless, extending the idea to our content is not trivial. Specifically, there are several key questions needed to be answered for validating this heuristic procedure:

- Does $\hat{k}^* \xrightarrow{p} k^*$, and if so, what is the convergence rate.
- Does $(\widehat{M}_p(\hat{\theta})/\hat{\sigma}_{\theta_1,p})(\hat{k}^*) \xrightarrow{p} (M_p^2(\theta_1)/\sigma_{\theta_1,p})(k^*)$, and if so, what is the convergence rate.

We leave these questions in future research.

Remark. On the other hand, Theorem 8 indicates that our test statistic might not be consistent against all fixed alternative hypotheses if $\varepsilon(Z;\theta_0)$ is collinear to the function $g(Z;\theta_0)$. However, given the nonlinear nature of our model, we do not think this type of alternatives.

We now proceed to consider the asymptotic local power properties. To this end, we study the asymptotic distribution of $n\widehat{\mathbb{M}}_p^2(\hat{\theta})$ under a certain sequence of Pitman-type local alternatives converging to null at a parametric rate:

$$H_{1,n} : \mathbb{E}\left(Y|X=x\right) = \mathcal{M}(x;\theta_0) + \frac{R(x)}{\sqrt{n}} \tag{18}$$

where the random variable $R(X)$ is P_X-integrable with zero mean and satisfies $P(R(X) = 0) < 1$.

Theorem 9 *Assume that* $\mathbb{M}_p^2(\theta) < \infty$ *for all* $\theta \in \Theta$*. Under* $H_{1,n}$*, we have*

$$\begin{aligned} n\widehat{\mathbb{M}}_n^2(\hat{\theta}) \xrightarrow{d} & \sum_{k=1}^{\infty} \tau_k^{(p)}(W_k^2 - 1) + 2N\left(0, 4\mathrm{Var}_{X,Z}\left(\mathbb{E}_{X',Z'}\varepsilon(Z;\theta_0)k_p(X,X')R(X')\right)\right) \\ & + \mathbb{E}\left(R(X)k_p(X,X')R(X')\right) \end{aligned} \tag{19}$$

where $\sum_{k=1}^{\infty}\tau_k^{(p)}(W_k^2-1)$ *is defined in Theorem 7.*

Proof See Appendix C. ■

Remark. A pathological situation in which our test will only have trivial local power against such alternatives is when $R(X)$ is a linear combination of $g(Z;\theta_0)$, i.e., $R(x) = \nu^\top g(Z;\theta_0)$ a.s. for some nonzero vector ν. In such a case, the limiting distribution of $n\widehat{\mathbb{M}}_p^2(\hat{\theta})$ under H_0 and $H_{1,n}$ is the same so that $H_{1,n}$ can not be detected. However, such a specific class of local alternatives is of very limited practical interest.

The following lemma states that the proposed test only have non-trivial local power in a finite-dimensional space, and there is only one direction with the highest asymptotic local power. Although this lemma is essentially Theorem 1 of Escanciano (2009), it provides a clear viewpoint that highlights the importance of a kernel.

To begin with, let T_k be an integral operator defined as

$$T_k f(x) = \int_{\mathcal{X}} k(x,x')f(x')dP_X(x')$$

Mercer's theorem states that one can characterize a kernel $k_p(\cdot,\cdot)$ as:

$$k_p(x,x') = \sum_{j\geq 1}\lambda_j e_j(x)e_j(x')$$

where the convergence is absolute and uniform, and $\{\lambda_j\}_{j\geq 1}, \{e_j(\cdot)\}_{j\geq 1}$ are eigenvalues and eigenfunctions of the operator T_k, respectively. $\lambda_1 > \lambda_2 > \cdots$ and $\lambda_j \to 0$ as $j\to\infty$. Since $\{e_j(\cdot)\}_{j\geq 1}$ are also basis of the space $L_2(\mathbb{R}^d, P_X)$, one can write $R(x) = \sum_{s\geq 1}\alpha_s e_s(x), \alpha_s = \langle R, e_s\rangle_{L_2(\mathbb{R}^d,P_X)} \in \mathbb{R}$.

Lemma 10 *Under local alternatives, let*

$$\mathbb{M}_p^2(\theta_0) = \mathbb{E}\left(\left(\varepsilon(Z;\theta_0) + \frac{R(X)}{\sqrt{n}}\right) k_p(X, X') \left(\varepsilon(Z';\theta_0) + \frac{R(X')}{\sqrt{n}}\right)\right)$$

we have

$$\mathbb{M}_p^2(\theta_0) = \mathbb{E}\left(\varepsilon(Z;\theta_0)k_p(X, X')\varepsilon(Z';\theta_0)\right) + \lambda_j \frac{\alpha_j^2}{n} + 2\lambda_j \mathbb{E}\left(\varepsilon(Z;\theta_0)e_j(X)\right) \frac{\alpha_j}{\sqrt{n}} \tag{20}$$

Proof See Appendix C. ■

Immediately, we conclude that if $\alpha_j \neq 0$ but $\{\alpha_s\}_{s\neq j} = 0$, then when $j = 1$, i.e., $R(x) = \alpha_1 e_1(x)$, the proposed test have highest asymptotic local power. The local power decreases when j increases, and when $j \to \infty$ one can only have trivial power. Nevertheless, for any fixed direction, e.g., $R(x) = \alpha_s e_s(x)$, we can change the value of λ_s (equivalently, change the kernel) to increase the local power.

4. A Multiplier Bootstrap Procedure

Our test statistic $n\widehat{\mathbb{M}}_p^2(\hat{\theta})$ is non-pivotal, in this section, we propose a simple-to-use multiplier bootstrap procedure to approximate the null distribution. Its implementation is listed below:

1. Generate a sequence of i.i.d random variables $\{v_i : i = 1, 2, \ldots, n\}$ with mean zero and variance one; e.g., Rademacher random variable, standard normal random variable, or Bernoulli random variable with $P(v = 1 - \kappa) = \kappa/\sqrt{5}$ and $P(v = \kappa) = 1 - \kappa/\sqrt{5}$, where $\kappa = (\sqrt{5} + 1)/2$ (Mammen, 1993).

2. Compute
$$\left(n\widehat{\mathbb{M}}_p^{2,*}(\hat{\theta})\right)_b = \frac{1}{n-1}\sum_{i\neq j} \varepsilon(z_i;\hat{\theta})v_i \hat{k}_p(x_i, x_j)\varepsilon(z_j;\hat{\theta})v_j$$

3. Repeat steps 1 and 2 B times, and collect $\left\{\left(n\widehat{\mathbb{M}}_p^{2,*}(\hat{\theta})\right)_b, b = 1, 2\ldots, B\right\}$

4. Define a confidence level α, obtain the $(1-\alpha)$-th quantile of $\left\{\left(n\widehat{\mathbb{M}}_p^{2,*}(\hat{\theta})\right)_b, b = 1, 2\ldots, B\right\}$, $c^*_{n,\alpha}$.

5. Reject the null if $n\widehat{\mathbb{M}}_p^2(\hat{\theta}) > c^*_{n,\alpha}$, and fail to reject otherwise.

The multiplier bootstrapped test statistic $n\widehat{\mathbb{M}}_p^{2,*}(\hat{\theta})$ has several attractive properties. First, it does not require computing new parameter estimates at each bootstrap draw, reducing the computational intensity of the proposed procedure. Second, due to the employment of the projection, its implementation does not require using estimators that admit an asymptotic linear representation. These computational conveniences are important when the dimension d is high.

The next theorem establishes the asymptotic validity of the proposed multiplier bootstrap procedure.

Theorem 11 *Assume that* $\mathbb{M}_p^2(\theta) < \infty$ *for all* $\theta \in \Theta$. *Then, we have* $n\widehat{\mathbb{M}}_p^{2,*}(\hat{\theta}) \overset{d,*}{\to} \sum_{k=1}^{\infty} \tau_k^{(p)}(W_k^2 - 1)$, *with probability one under the bootstrap law. Here* $\sum_{k=1}^{\infty} \tau_k^{(p)}(W_k^2 - 1)$ *is defined as the same in Theorem 7, and* $\overset{d,*}{\to}$ *denotes weak convergence under the bootstrap law, i.e., conditional on the original sample* $\{z_i, x_i : i = 1, 2, \ldots, n\}$.

Proof See Appendix C. ■

Theorem 11 states that the bootstrap statistic $n\widehat{\mathbb{M}}_p^{2,*}(\hat{\theta})$ converges to the null distribution of $n\widehat{\mathbb{M}}_p^2(\hat{\theta})$ conditional on the original sample under H_0, H_1 and $H_{1,n}$. This fact is what allows the proposed procedure to work.

5. A Minimum Distance Estimator

Based on the U-statistic expression derived in Section 2, we present a minimum distance estimator in this section. It is known that when the number of arbitrarily chosen instrument is finite, the GMM estimation procedure could render inconsistent estimates due to an identification problem, see, e.g., Domínguez and Lobato (2004) for various examples. Integrated conditional moment, on the other hand, introduces infinite many instruments, and therefore, does not arise the identification issue. Domínguez and Lobato (2004) is the first in the literature to introduce a consistent estimation procedure based on ICM framework, their estimator reads as,

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{n^3} \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \varepsilon(z_i;\theta)\mathbb{I}\{x_i \le x_j\} \right)^2$$

This estimator corresponds to an indicator weighting function $\mathbb{I}\{x \le u\}$, and suffers from the curse of dimensionality due to data sparseness.

The representation of ICM statistic in the RKHS provides a natural channel to develop a minimum estimator:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{n(n-1)} \sum_{i\neq j} \varepsilon(z_i;\theta)k(x_i, x_j)\varepsilon(z_j;\theta) \tag{21}$$

The objective function $\widehat{R}_U(\theta)$ can be rewritten as

$$\widehat{R}_U(\theta) = \varepsilon(z;\theta)^\top W_U \varepsilon(z;\theta)$$

where $W_U \in \mathbb{R}^{n\times n}$ is a symmetric weight matrix that depends on the kernel matrix K with $K_{i,j} = k(x_i, x_j)$. Here $W_U = (K - diag(K_{11}, \ldots, K_{nn}))/(n(n-1))$, where $diag(a_1, \ldots, a_n)$ denotes an $n \times n$ diagonal matrix whose diagonal elements are $a_1, \ldots, a_n$.

Although using the objective function $\widehat{R}_U(\theta)$, one could obtain a minimum-variance unbiased estimator, the weight matrix W_U, unfortunately, is indefinite, since $trace(W_U) = \sum_{i=1}^n \varpi_i = 0$, where $\{\varpi_i; i = 1, \ldots, n\}$ are the eigenvalues of W_U. Thus, we conclude that there exist both positive and negative eigenvalues.

We, therefore, focus on the V-statistic version of this estimator:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{n^2} \sum_{i,j=1}^{n} \varepsilon(z_i;\theta)k(x_i, x_j)\varepsilon(z_j;\theta) \tag{22}$$

whose objective function $\widehat{R}_V(\theta)$ can be written as

$$\widehat{R}_V(\theta) = \varepsilon(z;\theta)^\top W_V \varepsilon(z;\theta)$$

with $W_V = K/n^2$.

Let $R_k(\theta) = \mathbb{E}\left(f_\theta(V, V')\right)$, where $f_\theta(v, v') = \varepsilon(z;\theta)k(x, x')\varepsilon(z';\theta)$ with $v = (x, z)$, and denote $\|\cdot\|_F$ as the Frobenius norm. The following theorems establish the asymptotic properties of this estimator.

Theorem 12 *Assume that* $\mathbb{E}\left(|Y|^2 < \infty\right)$, $\mathbb{E}\left(\sup_{\theta\in\Theta}|\mathcal{M}(X;\theta)|^2\right) < \infty$, Θ *is compact and convex,* $R_k(\theta)$ *is uniquely minimized at* $\theta_0 \in \Theta^\circ$*, and Assumption A4 holds, then*

$$\hat{\theta} \xrightarrow{p} \theta_0$$

Proof See Appendix C. ■

Theorem 13 *Suppose that* $\mathcal{M}(X;\theta)$ *is twice continuously differentiable about* θ*,* Θ *is compact,* $H = \mathbb{E}\left(\nabla^2_\theta f_{\theta_0}(V, V')\right)$ *is non-singular,* $\mathbb{E}\left(|Y|^2 < \infty\right)$, $\mathbb{E}\left(\sup_{\theta\in\Theta}|\mathcal{M}(X;\theta)|^2\right) < \infty$, $\mathbb{E}\left(\sup_{\theta\in\Theta}\|\nabla_\theta\mathcal{M}(X;\theta)\|_2^2\right) < \infty$, $\mathbb{E}\left(\sup_{\theta\in\Theta}\|\nabla^2_\theta\mathcal{M}(X;\theta)\|_F^2\right) < \infty$, $R_k(\theta)$ *is uniquely minimized at* $\theta_0 \in \Theta^\circ$*, and Assumption A4 holds, then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma_V)$$

where

$$\Sigma_V = 4H^{-1}\mathrm{Var}_V\left(\mathbb{E}^2_{V'}(\nabla_\theta f_{\theta_0}(V, V'))\right)H^{-1}$$

Proof See Appendix C. ■

In general, the estimator given by (22) is not efficient. An efficient estimator based on infinite number of moment conditions can be constructed following the ideas of Carrasco and Florens (2000). For regularized and infinite dimensional estimators based on $\widehat{R}_{U(V)}(\theta)$, see Zhang et al. (2020).

6. Monte Carlo Studies

This section conducts a sequence of Monte Carlo simulations to evaluate the finite sample performance of the kernel-based tests. Before specifying the data generating processes (DGPs), first observe that for any $\theta \in \Theta$, we have

$$\mathbb{E}\left[\varepsilon(Z;\theta)k_p(X,X')\varepsilon(Z';\theta)\right] = \mathbb{E}\left[\varepsilon_p(Z;\theta)k(X,X')\varepsilon_p(Z';\theta)\right] \tag{23}$$

where

$$\varepsilon_p(z;\theta) = \varepsilon(z;\theta) - g^{\top}(z;\theta)\Gamma_{\theta_0}^{-1}\mathbb{E}\left(g(Z;\theta)\varepsilon(Z;\theta)\right)$$

is the projection residual of $\varepsilon(z;\theta)$. Its verification can be found in Appendix C. This representation greatly simplifies computation, as

$$n\widehat{\mathbb{M}}_p^2(\hat{\theta}) = \frac{1}{n-1}\sum_{i\neq j}\hat{\varepsilon}_p(z_i;\hat{\theta})k(x_i,x_j)\hat{\varepsilon}_p(z_j;\hat{\theta})$$

and

$$n\widehat{\mathbb{M}}_p^{2,*}(\hat{\theta}) = \frac{1}{n-1}\sum_{i\neq j}\hat{\varepsilon}_p(z_i;\hat{\theta})v_i k(x_i,x_j)\hat{\varepsilon}_p(z_j;\hat{\theta})v_j$$

where

$$\hat{\varepsilon}_p(z_i;\hat{\theta}) = \varepsilon(z_i;\hat{\theta}) - g^{\top}(z_i)\Gamma_{n,\hat{\theta}}^{-1}\left(\frac{1}{n}\sum_{s=1}^{n}g(z_s;\hat{\theta})\varepsilon(z_s;\hat{\theta})\right)$$

and $\Gamma_{n,\hat{\theta}}$ is the empirical counterpart of Γ_{θ_0}.

6.1 Data Generating Processes

We consider the following DGPs:

- DGP(m): $Y_i = \beta_0 + \sum_{j=1}^{m}\beta_j X_{ji} + \sigma_i^{(m)}\varepsilon_i$.
- DGP-LOCAL(m): $Y_i = \beta_0 + \sum_{j=1}^{m}\beta_j X_{ji} + n^{-1/2}\sum_{j=1}^{m}\beta_j X_{ji}^2 + \sigma_i^{(m)}\varepsilon_i$.

DGP(m) specifies m covariates and is used to evaluate the size performance of proposed tests. DGP-LOCAL(m) is used to evaluate local powers of the corresponding null DGPs.

We allow for conditional heteroskedasticity in all models and generate the covariates and heteroskedasticity as follows.

In DGP(m) and DGP-LOCAL(m),

- When $m = 2$, $X_1, X_2 \sim N(0,1)$, and $\sigma^{(2)} = \left(0.1 + X_1^2 + X_2^2\right)^{1/2}$.
- When $m = 5$, $X_j \sim U(0,j)$ for $j = 1,2,3$, $X_j \sim N(0,(j-3)^2)$ for $j = 4,5$. $\sigma^{(5)} = \left(0.1 + \sum_{j=1}^{3}X_j + \sum_{j=4}^{5}X_j^2\right)^{1/2}$.

- When $m = 10$, $X_j \sim U(0, j)$ for $j = 1, \ldots, 5$, $X_j \sim N(0, (j-5)^2)$ for $j = 6, \ldots, 10$. $\sigma^{(10)} = \left(0.1 + \sum_{j=1}^{5} X_j + \sum_{j=5}^{10} X_j^2\right)^{1/2}$.
- When $m = 20$, $X_j \sim U(0, j)$ for $j = 1, \ldots, 10$, $X_j \sim N(0, (j-10)^2)$ for $j = 11, \ldots, 20$. $\sigma^{(20)} = \left(0.1 + \sum_{j=1}^{10} X_j + \sum_{j=11}^{20} X_j^2\right)^{1/2}$.

In call cases, we set $\varepsilon_i \sim N(0, 1)$, and set β_j's to be 1.

6.2 Test Statistics and Simulation Results

From the discussion of asymptotic power properties, it should be clear now that the choice of a kernel is important for ICM testing. Nevertheless, finding a case-dependent optimal kernel is challenging. In this subsection, we provide a heuristic algorithm for tunning the parameter of a Gaussian kernel $k(x, x') = \exp(-\gamma \|x - x'\|_2^2), \gamma > 0$. Note that we are not claiming such algorithm would lead to an optimal testing statistic. Rather, we believe that this heuristic algorithm would not deliver a bad test (in the sense of small power) with greater probability.

The Gaussian kernel is the default kernel in many kernel-based algorithms. Conventional wisdom suggest that 'Gaussian kernels tend to yield good performance under general smoothness assumptions and should be considered especially if no additional knowledge of the data is available' (Smola et al., 1998). A Gaussian kernel takes the form of a normal distribution and is smooth. The tunning parameter γ determines how well this kernel fit the data $\mathbb{X}$, here $\mathbb{X}$ is a $n \times d$ matrix consisting of conditional variables. Fixing an input data x', a large γ would lead to an over-fitting scenario, since large weight would concentrated around x', while points that are far away from x' would have a kernel value that decay to zero exponentially. A small γ corresponds to an under-fitting scenario, as points would have kernel values close to one. These two cases are essentially the same thing: the resulting kernel values concentrate around one point (zero or one), making a test powerless.

The desired parameter γ would 'spread' the kernel values in the range $(0, 1]$, one nature method is to normalize the data using the second moment information of the input matrix $\mathbb{X}$. We propose to perform a principal component analysis (PCA) for $\mathbb{X}$, and set $\gamma = 1/(2\zeta_1)$, where ζ_1 is the largest principal value. The idea is simple: In high-dimensional cases, the norm $\|x - x'\|_2^2 = \sum_{s=1}^{d} |x_s - x'_s|^2$ are more likely to make kernel values concentrate around zero than one, and the higher the variance of the data, the higher the probability of occurring such concentration. By setting $\gamma = 1/(2\zeta_1)$, one could avoid such phenomenon.

We consider five kernels in simulation studies:

- Guassian Kernel, $k_1(x, x') = \exp(-(1/\zeta_1)\|x - x'\|_2^2)$.
- Inverse Multiquadric (IMQ) Kernel, $k_2(x, x') = \left(1 + \|x - x'\|_2^2\right)^{-1.5}$.
- Gaussian+IMQ Kernel, $k_3(x, x') = k_1(x, x') + \left(1 + \|x - x'\|_2^2\right)^{-0.5}$.

- Shift Variant Kernel, $k_4(x, x') = (2 + \sin(4\|x\|_2))k_1(x, x')(2 + \sin(4\|x'\|_2))$
- Local Periodic Kernel, $k_5(x, x') = (2 + \sin(0.1\|x - x'\|_2^2))k_1(x, x')$.

The Gaussian+IMQ kernel, the Shift Variant kernel and the Local Periodic kernel are constructed as results of Lemmas 2, 3 and 4, respectively.

We report the simulation results in Table 2. The nominal significance levels are given by $0.01, 0.05$ and 0.1, while the sample sizes range from $N = 100, N = 200$ to $N = 400$. For each experiment, i.e., each DGP and sample size, we run 1000 simulations. For each round of simulation, the bootstrap procedure repeats 500 times to estimate the critical value. The parameters β_j's are estimated by ordinary least-squares.

Table 2: Simulation Results, Five Kernels

N=100	0.1					0.05					0.01				
DGP	Gaussian	IMQ	Gaussian+IMQ	Shift Variant	Local Periodic	Gaussian	IMQ	Gaussian+IMQ	Shift Variant	Local Periodic	Gaussian	IMQ	Gaussian+IMQ	Shift Variant	Local Periodic
DGP(2)	0.113	0.124	0.113	0.116	0.118	0.068	0.063	0.065	0.065	0.065	0.011	0.016	0.017	0.015	0.016
DGP(5)	0.112	0.109	0.107	0.101	0.102	0.054	0.057	0.059	0.051	0.059	0.009	0.013	0.017	0.01	0.014
DGP(10)	0.109	0.117	0.116	0.109	0.109	0.055	0.057	0.062	0.054	0.059	0.015	0.007	0.02	0.02	0.018
DGP(20)	0.081	0.081	0.082	0.102	0.11	0.053	0.047	0.047	0.064	0.07	0.016	0.011	0.02	0.02	0.019
DGP-LOCAL(2)	0.247	0.205	0.213	0.255	0.238	0.17	0.129	0.132	0.179	0.155	0.057	0.045	0.04	0.052	0.045
DGP-LOCAL(5)	0.329	0.293	0.322	0.297	0.348	0.247	0.205	0.24	0.21	0.233	0.098	0.086	0.102	0.064	0.101
DGP-LOCAL(10)	0.963	0.907	0.965	0.939	0.948	0.938	0.86	0.946	0.888	0.923	0.86	0.718	0.849	0.796	0.821
DGP-LOCAL(20)	0.999	1	1	1	1	0.999	1	1	0.998	1	0.999	1	0.998	0.992	0.999
N=200	0.1					0.05					0.01				
DGP	Gaussian	IMQ	Gaussian+IMQ	Shift Variant	Local Periodic	Gaussian	IMQ	Gaussian+IMQ	Shift Variant	Local Periodic	Gaussian	IMQ	Gaussian+IMQ	Shift Variant	Local Periodic
DGP(2)	0.111	0.113	0.111	0.109	0.123	0.053	0.059	0.056	0.046	0.065	0.007	0.013	0.009	0.011	0.013
DGP(5)	0.101	0.127	0.103	0.105	0.111	0.05	0.062	0.046	0.058	0.064	0.01	0.01	0.01	0.012	0.009
DGP(10)	0.099	0.112	0.097	0.119	0.154	0.054	0.058	0.044	0.069	0.091	0.012	0.016	0.011	0.019	0.023
DGP(20)	0.108	0.093	0.109	0.116	0.112	0.055	0.05	0.059	0.061	0.062	0.013	0.012	0.015	0.015	0.015
DGP-LOCAL(2)	0.234	0.228	0.212	0.234	0.239	0.149	0.134	0.137	0.153	0.146	0.045	0.045	0.047	0.052	0.038
DGP-LOCAL(5)	0.347	0.284	0.352	0.314	0.353	0.261	0.178	0.272	0.218	0.264	0.108	0.065	0.119	0.085	0.115
DGP-LOCAL(10)	0.972	0.931	0.977	0.962	0.977	0.958	0.903	0.958	0.933	0.953	0.891	0.778	0.878	0.81	0.855
DGP-LOCAL(20)	1	1	1	1	1	1	1	1	1	1	1	1	1	0.999	1
N=400	0.1					0.05					0.01				
DGP	Gaussian	IMQ	Gaussian+IMQ	Shift Variant	Local Periodic	Gaussian	IMQ	Gaussian+IMQ	Shift Variant	Local Periodic	Gaussian	IMQ	Gaussian+IMQ	Shift Variant	Local Periodic
DGP(2)	0.101	0.119	0.099	0.098	0.126	0.051	0.066	0.044	0.047	0.066	0.013	0.007	0.009	0.01	0.022
DGP(5)	0.1	0.098	0.102	0.105	0.119	0.051	0.057	0.05	0.051	0.061	0.011	0.008	0.016	0.013	0.011
DGP(10)	0.102	0.124	0.106	0.118	0.094	0.049	0.062	0.061	0.064	0.051	0.014	0.013	0.015	0.019	0.012
DGP(20)	0.12	0.095	0.118	0.102	0.109	0.068	0.051	0.054	0.059	0.058	0.013	0.011	0.009	0.015	0.011
DGP-LOCAL(2)	0.214	0.193	0.205	0.227	0.225	0.137	0.119	0.117	0.152	0.145	0.05	0.04	0.04	0.044	0.057
DGP-LOCAL(5)	0.338	0.294	0.353	0.3	0.365	0.239	0.192	0.236	0.208	0.278	0.103	0.068	0.091	0.072	0.114
DGP-LOCAL(10)	0.981	0.955	0.988	0.959	0.977	0.97	0.926	0.974	0.933	0.959	0.906	0.816	0.905	0.824	0.892
DGP-LOCAL(20)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

We first analyze the size of proposed tests. From the results of $DGP(m), m = 2, 5, 10, 20$, we find that the actual finite sample size of all the proposed tests are close to their nominal size, even when the sample size is as small as 100 and the dimension is as high as 20. Results of $DGP\text{-}LOCAL(m), m = 2, 5, 10, 20$ confirm that all proposed tests have non-trivial local power, the power is especially high when tests are facing high-dimensional models. We want to emphasize that the size as well as the power of conventional CMR tests often diminish rapidly to zero as dimension increases, and the degeneracy of levels or powers does not improve when sample size increases. Lastly, we want to emphasize that different kernels (different test statistics) have different power against different models. Each test necessarily exploits certain features of the data generating process at the expense of others, complementarities between tests can easily arise.

To further illustrate our intuition on γ, we propose a parallel set of DGPs and focus on Bierens' test statistic, which is a V-statistic associated with a Gaussian kernel with tunning parameter $\gamma = 1/2$:

$$T_{n,v} = \frac{1}{n} \sum_{j,k=1}^{n} \varepsilon(z_j; \hat{\theta}) \varepsilon(z_k; \hat{\theta}) \exp\left(-\frac{1}{2} \|x_j - x_k\|^2\right)$$

its U-statistic version is

$$T_{n,u} = \frac{1}{n-1} \sum_{j \neq k} \varepsilon(z_j; \hat{\theta}) \varepsilon(z_k; \hat{\theta}) \exp\left(-\frac{1}{2} \|x_j - x_k\|^2\right)$$

The set of parallel DGPs are identical to previous ones (and denoted as DGP(m)* and DGP-LOCAL(m)*), except in the following areas,

- When $m = 2$, $X_1, X_2 \sim N(0, 10)$, and $\sigma^{(2)} = \left(0.1 + X_1^2 + X_2^2\right)^{1/2}$.
- When $m = 5$, $X_j \sim U(0, 10 * j)$ for $j = 1, 2, 3$, $X_j \sim N(0, (10 * (j-3))^2)$ for $j = 4, 5$. $\sigma^{(5)} = \left(0.1 + \sum_{j=1}^{3} X_j + \sum_{j=4}^{5} X_j^2\right)^{1/2}$.
- When $m = 10$, $X_j \sim U(0, 1 + 0.1 * (j-1))$ for $j = 1, \ldots, 5$, $X_j \sim N(0, (1 + 0.1 * (j-5))^2)$ for $j = 6, \ldots, 10$. $\sigma^{(10)} = \left(0.1 + \sum_{j=1}^{5} X_j + \sum_{j=5}^{10} X_j^2\right)^{1/2}$.
- When $m = 20$, $X_j \sim U(0, 1 + 0.1 * (j-1))$ for $j = 1, \ldots, 10$, $X_j \sim N(0, (1 + 0.1 * (j-11))^2)$ for $j = 11, \ldots, 15$, $X_j \sim N(1, (1 + 0.1 * (j-15))^2)$ for $j = 15, \ldots, 20$. $\sigma^{(20)} = \left(0.1 + \sum_{j=1}^{10} X_j + \sum_{j=11}^{20} X_j^2\right)^{1/2}$.

In a nutshell, this set of parallel DGPs only differ in the variances of conditional variables such that in low-dimensional cases (i.e., DGP(2)*,DGP(5)*,DGP-LOCAL(2)*,DGP-LOCAL(5)*), the largest principal value $\zeta_1 > 2$, and in high-dimensional cases (i.e., rest of the DGPs), the largest principal value $\zeta_1 \in (1, 2.25)$. While in contrast, in the original DGPs, we have $\zeta_1 \in (1, 2.25)$ for low-dimensional cases and $\zeta_1 > 2$ for high-dimensional cases.

Tables 3-5 present the results. We also study the performance of a Gaussian kernel with tunning parameter $\gamma = 1/(2\zeta_1)$ under these parallel DGPs, the results are shown in the Table 6. We draw the following remarks:

- Overall, V-Statistics are inferior to U-statistics, this is especially true when conditional variables have high variance, i.e., the parallel DGPs. At extreme cases, the V-Statistic test sizes are completely wrong, lost almost all the power against alternatives. In the contrast, U-Statistic perform well when variance of conditional variables match the tunning parameter (e.g., low-dimensional cases in the original DGPs and high-dimensional cases in the parallel DGPs). We conjecture that this is because V-statistic are biased statistic, and high dimension increases the bias level.

- Tunning parameter works as conjectured. High-dimensional cases in the parallel DGPs have accurate size and good power against local alternatives,but these cases perform poorly under the original DGPs. A reverse pattern holds true for low-dimensional cases under parallel and original cases.
- Model complexity (i.e., dimension) is important. Observe that even though in the parallel DGPs, low-dimensional cases are mis-matched by its tunning parameter, DGP(2)* has more accurate test size compare to DGP(5)*, and DGP-LOCAL(2)* has more power than DGP-LOCAL(5)*. These pattern suggest model complexity plays an important role in a test's power properties.

Table 3: Bierens' Test with significant level $\alpha = 0.1$

$\alpha = 0.1$	U-Statistic			V-Statistic		
	N=100	N=200	N=400	N=100	N=200	N=400
DGP(2)	0.116	0.126	0.105	0.148	0.111	0.124
DGP(5)	0.12	0.11	0.095	0.1	0.081	0.091
DGP(10)	0.127	0.127	0.117	0	0	0
DGP(20)	0.188	0.175	0.141	0.699	0	0
DGP-LOCAL(2)	0.226	0.222	0.2	0.224	0.209	0.212
DGP-LOCAL(5)	0.289	0.312	0.277	0.227	0.264	0.294
DGP-LOCAL(10)	0.378	0.398	0.371	0	0	0
DGP-LOCAL(20)	0.232	0.207	0.199	0.694	0	0
$\alpha = 0.1$	U-Statistic			V-Statistic		
	N=100	N=200	N=400	N=100	N=200	N=400
DGP(2)*	0.104	0.11	0.097	0.001	0	0
DGP(5)*	0.092	0.126	0.158	0	0	0
DGP(10)*	0.134	0.116	0.116	0.027	0.01	0.022
DGP(20)*	0.129	0.117	0.113	0.731	0	0
DGP-LOCAL(2)*	0.996	1	1	0.121	0.616	0.958
DGP-LOCAL(5)*	0.41	0.581	0.713	0	0	0
DGP-LOCAL(10)*	0.919	0.954	0.973	0.692	0.795	0.881
DGP-LOCAL(20)*	0.505	0.589	0.611	0.812	0	0

7. An Empirical Illustration

This section demonstrates the proposed method with an empirical application. This application studies a basic income-determination model between house household expenditure (Y) and income (X). The data used in this section comes from Berger (2022), and can be downloaded from (https://doi.org/10.1093/ectj/utab032). The data is a simulated version of the 1998-99 UK Family Expenditure Survey (FES)[3]. The FES is a random sample stratified by regions, it contains $n = 6630$ households drawn from the Post Office's list of addresses.

3. Due to legal reasons, Berger (2022) is unable to provide actual FES data as part of his paper. Instead, a simulated data, similar to the actual data, are provided.

Table 4: Bierens' Test with significant level $\alpha = 0.05$

$\alpha = 0.05$	U-Statistic			V-Statistic		
	N=100	N=200	N=400	N=100	N=200	N=400
DGP(2)	0.062	0.065	0.056	0.071	0.049	0.065
DGP(5)	0.068	0.056	0.043	0.034	0.037	0.04
DGP(10)	0.034	0.033	0.036	0	0	0
DGP(20)	0.137	0.085	0.035	0.01	0	0
DGP-LOCAL(2)	0.144	0.142	0.12	0.122	0.121	0.15
DGP-LOCAL(5)	0.196	0.215	0.185	0.113	0.166	0.168
DGP-LOCAL(10)	0.168	0.214	0.224	0	0	0
DGP-LOCAL(20)	0.152	0.083	0.041	0.014	0	0
$\alpha = 0.05$	U-Statistic			V-Statistic		
	N=100	N=200	N=400	N=100	N=200	N=400
DGP(2)*	0.041	0.057	0.044	0	0	0
DGP(5)*	0.002	0.001	0.009	0	0	0
DGP(10)*	0.056	0.058	0.069	0.001	0	0.003
DGP(20)*	0.074	0.056	0.044	0.009	0	0
DGP-LOCAL(2)*	0.988	0.999	1	0.021	0.285	0.855
DGP-LOCAL(5)*	0.016	0.039	0.183	0	0	0
DGP-LOCAL(10)*	0.88	0.922	0.955	0.335	0.583	0.749
DGP-LOCAL(20)*	0.403	0.453	0.492	0.018	0	0

Table 5: Bierens' Test with significant level $\alpha = 0.01$

$\alpha = 0.01$	U-Statistic			V-Statistic		
	N=100	N=200	N=400	N=100	N=200	N=400
DGP(2)	0.011	0.014	0.01	0.012	0.011	0.011
DGP(5)	0.016	0.013	0.007	0.005	0.007	0.006
DGP(10)	0	0	0.006	0	0	0
DGP(20)	0.044	0.018	0.011	0	0	0
DGP-LOCAL(2)	0.048	0.05	0.033	0.028	0.034	0.047
DGP-LOCAL(5)	0.074	0.075	0.076	0.025	0.039	0.047
DGP-LOCAL(10)	0.012	0.036	0.055	0	0	0
DGP-LOCAL(20)	0.042	0.021	0.01	0	0	0
$\alpha = 0.01$	U-Statistic			V-Statistic		
	N=100	N=200	N=400	N=100	N=200	N=400
DGP(2)*	0.003	0.007	0.006	0	0	0
DGP(5)*	0	0	0	0	0	0
DGP(10)*	0.012	0.01	0.012	0	0	0
DGP(20)*	0.011	0.007	0.006	0	0	0
DGP-LOCAL(2)*	0.874	0.984	1	0	0.018	0.428
DGP-LOCAL(5)*	0	0	0.005	0	0	0
DGP-LOCAL(10)*	0.706	0.836	0.881	0.039	0.181	0.417
DGP-LOCAL(20)*	0.193	0.205	0.263	0	0	0

The model specifies as

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The income X is potentially endogenous, and following Berger (2022), an instrument variable $Z = X - Y$ is created. The null hypothesis is that Z is a valid instrument variable,

Table 6: Gaussian Kernel with $\gamma = 1/(2\zeta_1)$ Under Parallel DGPs

	$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400
DGP(2)*	0.141	0.106	0.104	0.067	0.054	0.054	0.022	0.013	0.015
DGP(5)*	0.129	0.094	0.116	0.072	0.044	0.06	0.016	0.015	0.015
DGP(10)*	0.107	0.101	0.098	0.052	0.058	0.052	0.016	0.014	0.009
DGP(20)*	0.089	0.093	0.116	0.062	0.043	0.055	0.019	0.012	0.013
DGP-LOCAL(2)*	1	1	1	1	1	1	1	1	1
DGP-LOCAL(5)*	1	1	1	1	1	1	1	1	1
DGP-LOCAL(10)*	0.979	0.992	0.992	0.965	0.977	0.988	0.903	0.935	0.958
DGP-LOCAL(20)*	0.788	0.871	0.891	0.731	0.802	0.836	0.592	0.651	0.695

i.e.,

$$H_0 : \mathbb{E}\left(\varepsilon \mid Z\right) = 0$$

The alternative hypothesis is just the negation of the null. We estimate parameters using the classical 2SLS method, and choose the Guassian Kernel, $k(x, x') = \exp(-(2/d)^2\|x - x'\|_2^2)$, where d is the dimension, to perform the test.

In table 7 we report the $p-$values of the 12 regions. We found that in most regions, Z could be qualified as valid IV, but all the $p-$values are close to the 5% significant level.

Table 7: p-values for testing IV validity

Region	p-value	Region	p-value	Region	p-value	Region	p-value
1	0.086	4	0.038	7	0.082	10	0.076
2	0.066	5	0.066	8	0.12	11	0.094
3	0.074	6	0.044	9	0.07	12	0.09

8. Conclusion

In this paper, we propose to represent ICM tests in the RKHS. There are several motivations behind this representation. First, conventional ICM tests are based on empirical processes and require an integration to obtain the Cramer-Von Mises statistics. This integration is often unable to present a closed form test statistics. Applications of ICM tests are then forced to focus on Birens's or Escanciano's ICM tests, which enjoy closed form presentations. Second, existing literature has well documented that when conditional variables are of high dimensional, ICM tests typically have power-loss issues. Existing dimension-reduction tools rely on projecting the covariates onto a one-dimensional space, and integrate projected statistics from all directions. This procedure leads to a kernel that is hard to compute (its algorithm complexity is $O(n^3)$). Third, although ICM tests are admissible, i.e., there exist no test that is uniformly more powerful than ICM tests, they do not have non-trivial power in all directions. In fact, one ICM test only has substantial power in a finite-dimensional space (Escanciano, 2009). Thus, it is desired to have as many ICM tests as possible.

Once we represent ICM tests in RKHS, we found that (i) after specifying a kernel, the CvM statistics is a closed-form U-statistic; (ii) a kernel embodies both the dimension and integral measure, and hence, is a valid dimension reduction tool; (iii) with some assumptions, new kernels could be constructed using existing kernels by addition or multiplication.

The main idea behind this representation consists of several steps: (i) the conditional moment restriction is transformed into an unconditional moment restriction with infinite number of moment conditions. This transformation is over a function space, and under the null, the supremum (over this function space) value of the squared unconditional moment condition is zero, i.e., the maximum moment restriction; (ii) To obtain a closed form of the maximum moment restriction, one could restrict the function space to a unit ball of a RKHS $\mathcal{H}(k)$. The maximum moment restriction corresponds to an integral operator, and Riesz's representation theorem states that there exists a unique element in the corresponding $\mathcal{H}(k)$ that represents such an integral operator. We call such element in $\mathcal{H}(k)$ the conditional moment embedding, and its $\mathcal{H}(k)-$norm is the norm of the integral operation. Under the null, this norm should be zero; (iii) Using the kernel trick, the squared norm has a closed form presentation, and we can estimate it and further build a test statistic using a U-statistic.

Kernels are essential in this framework, only ISPD kernels could lead a conditional moment embedding that is injective to the original null hypothesis. Commonly used ISPD kernels are the Gaussian, the Laplacian, the Inverse multiquadric and the Matern kernels. One could also construct new ISPD kernels from existing ones with additional assumptions imposed.

We further propose a projected kernel to eliminate estimation effects. The advantages of using such a kernel are (i) the limiting null distribution of the test statistics does not depend on how an estimator is obtained; (ii) We do not need to require the estimator is $\sqrt{n}-$asymptotically linear; (iii) The corresponding tests could include certain 'non-standard' estimators whose convergence rate is slower than $1/\sqrt{n}$. We propose a simple multiplier bootstrap to find the critical value. This is particularly appreciated if the underlying model is non-linear and estimation is time-consuming.

A minimum distance estimator based on the conditional moment embedding is developed as a byproduct. This estimator inherits merits of the corresponding test statistic, e.g., dimension-reduction properties.

Monte Carlo experiments are conducted, and simulation results indicate that the proposed tests have accurate finite sample size and admirably well local power even when the sample size is as small as $n = 100$ and the dimension is as high as $d = 20$. In addition, simulation results also suggest that the proposed tests have good power against high-frequency alternatives. Lastly, a simple empirical application is studied.

References

Aronszajn, N. (1950): "Theory of reproducing kernels," *Transactions of the American mathematical society*, 68, 337–404.

Berger, Y. G. (2022): "Testing conditional moment restriction models using empirical likelihood," *The Econometrics Journal*, 25, 384–403.

Bickel, P. J., Y. Ritov, and T. M. Stoker (2006): "Tailor-made tests for goodness of fit to semiparametric hypotheses," *The Annals of Statistics*, 34, 721–741.

Bierens, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20, 105–134.

Bierens, H. J. and W. Ploberger (1997): "Asymptotic theory of integrated conditional moment tests," *Econometrica: Journal of the Econometric Society*, 1129–1151.

Bochner, S. (1933): "Monotone funktionen, stieltjessche integrale und harmonische analyse," *Mathematische Annalen*, 108, 378–410.

Carrasco, M. and J.-P. Florens (2000): "Generalization of GMM to a continuum of moment conditions," *Econometric Theory*, 16, 797–834.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization," *Handbook of econometrics*, 6, 5633–5751.

Delgado, M. A., M. A. Domínguez, and P. Lavergne (2006): "Consistent tests of conditional moment restrictions," *Annales d'Économie et de Statistique*, 33–67.

Delgado, M. A. and W. Stute (2008): "Distribution-free specification tests of conditional models," *Journal of Econometrics*, 143, 37–55.

Dinculeanu, N. (2000): *Vector integration and stochastic integration in Banach spaces*, vol. 48, John Wiley & Sons.

Domínguez, M. A. and I. N. Lobato (2004): "Consistent estimation of models defined by conditional moment restrictions," *Econometrica*, 72, 1601–1615.

——— (2015): "A simple omnibus overidentification specification test for time series econometric models," *Econometric Theory*, 31, 891–910.

Escanciano, J. C. (2006a): "A consistent diagnostic test for regression models using projections," *Econometric Theory*, 22, 1030–1051.

——— (2006b): "Goodness-of-fit tests for linear and nonlinear time series models," *Journal of the American Statistical Association*, 101, 531–541.

——— (2009): "On the lack of power of omnibus specification tests," *Econometric Theory*, 25, 162–194.

Escanciano, J. C. and S.-C. Goh (2014): "Specification analysis of linear quantile models," *Journal of Econometrics*, 178, 495–507.

Fan, Y. and Q. Li (2000): "Consistent model specification tests: Kernel-based tests versus Bierens' ICM tests," *Econometric Theory*, 16, 1016–1041.

Gretton, A., D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur (2012): "Optimal kernel choice for large-scale two-sample tests," *Advances in neural information processing systems*, 25.

Guo, X. and L. Zhu (2017): "A review on dimension-reduction based tests for regressions," *From statistics to mathematical finance:Festschrift in Honour of Winfried Stute*, 105–125.

Hofmann, T., B. Schölkopf, and A. J. Smola (2008): "Kernel methods in machine learning," *The annals of statistics*, 36, 1171–1220.

Khmaladze, E. V. (1982): "Martingale approach in the theory of goodness-of-fit tests," *Theory of Probability & Its Applications*, 26, 240–257.

——— (1993): "Goodness of fit problem and scanning innovation martingales," *The Annals of Statistics*, 798–829.

Kimeldorf, G. and G. Wahba (1971): "Some results on Tchebycheffian spline functions," *Journal of mathematical analysis and applications*, 33, 82–95.

Koul, H. L. and W. Stute (1999): "Nonparametric model checks for time series," *The Annals of Statistics*, 27, 204–236.

Lavergne, P. and V. Patilea (2012): "One for all and all for one: regression checks with many regressors," *Journal of business & economic statistics*, 30, 41–52.

Mammen, E. (1993): "Bootstrap and wild bootstrap for high dimensional linear models," *The annals of statistics*, 21, 255–285.

Minh, H. Q., P. Niyogi, and Y. Yao (2006): "Mercer's theorem, feature maps, and smoothing," in *International Conference on Computational Learning Theory*, Springer, 154–168.

Muandet, K., K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. (2017): "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, 10, 1–141.

Muandet, K., W. Jitkrittum, and J. Kübler (2020): "Kernel conditional moment test via maximum moment restriction," in *Conference on Uncertainty in Artificial Intelligence*, PMLR, 41–50.

Newey, W. K. (1985): "Generalized method of moments specification testing," *Journal of econometrics*, 29, 229–256.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Neyman, J. (1959): "Optimal tests of composite statistical hypotheses," *The Harald Cram& Volume*, 213–234.

Paulsen, V. I. and M. Raghupathi (2016): *An introduction to the theory of reproducing kernel Hilbert spaces*, vol. 152, Cambridge university press.

Robinson, P. M. (1991): "Best nonlinear three-stage least squares estimation of certain econometric models," *Econometrica: Journal of the Econometric Society*, 755–786.

Rudin, W. (2017): *Fourier analysis on groups*, Courier Dover Publications.

Sant'Anna, P. H. and X. Song (2020): "Specification tests for generalized propensity scores using double projections," *arXiv preprint arXiv:2003.13803*.

Sant'Anna, P. H. and X. Song (2019): "Specification tests for the propensity score," *Journal of Econometrics*, 210, 379–404.

Serfling, R. J. (1980): *Approximation theorems of mathematical statistics*, John Wiley & Sons.

Smola, A. J., B. Schölkopf, and K.-R. Müller (1998): "The connection between regularization operators and support vector kernels," *Neural networks*, 11, 637–649.

Sriperumbudur, B. K., A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet (2010): "Hilbert space embeddings and metrics on probability measures," *The Journal of Machine Learning Research*, 11, 1517–1561.

Steinwart, I. (2001): "On the influence of the kernel on the consistency of support vector machines," *Journal of machine learning research*, 2, 67–93.

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

Tolstikhin, I., B. K. Sriperumbudur, and K. Muandet (2017): "Minimax estimation of kernel mean embeddings," *The Journal of Machine Learning Research*, 18, 3002–3048.

Wendland, H. (2004): *Scattered data approximation*, vol. 17, Cambridge university press.

Zhang, R., M. Imaizumi, B. Schölkopf, and K. Muandet (2020): "Maximum moment restriction for instrumental variable regression," *arXiv preprint arXiv:2010.07684*.

A. Backgrounds on the Reproducing Kernel Hilbert Space

Reproducing Kernel Hilbert Space (RKHS), first proposed in Aronszajn (1950), is a special Hilbert space with some properties. It is a Hilbert space of functions with reproducing kernels. Formally, RKHS is defined as

Definition 14 *A Reproducing Kernel Hilbert Space is a Hilbert space $\mathcal{H}(k)$ of functions $f : \mathcal{X} \to \mathbb{R}$ with a reproducing kernel $k : \mathcal{X}^2 \to \mathbb{R}$, where (i) $k(x, \cdot) \in \mathcal{H}(k)$, and (ii) $\langle f, k(x, \cdot)\rangle_{\mathcal{H}(k)} = f(x)$*

Remark. Property (ii) is called the reproducing property of $\mathcal{H}(k)$. By Aronszajn (1950), every positive definite kernel k uniquely determines the RKHS for which k is a reproducing kernel.

To gain understanding about the kernel k, some backgrounds are needed.

Definition 15 *Given a kernel $k : \mathcal{X}^2 \to \mathbb{R}$ and inputs $x_1, \ldots, x_n \in \mathcal{X}$, the $n \times n$ matrix*

$$K_{ij} = k(x_i, x_j), \forall i, j \in \{1, \ldots, n\}$$

is called the Gram Matrix (also known as the Kernel Matrix) of k with respect to $x_1, \ldots, x_n$.

Definition 16 *A real $n \times n$ symmetric matrix K_{ij} satisfying*

$$\sum_{i,j=1}^{n} c_i c_j K_{ij} \geq 0$$

for all $c_i \in \mathbb{R}$ is called positive definite. If equality in the above equation only occurs $c_1 = \cdots = c_n = 0$, then we shall call the matrix strictly positive definite.

Definition 17 *Let $\mathcal{X}$ be a nonempty set. A function $k : \mathcal{X}^2 \to \mathbb{R}$ which for all $n \in \mathbb{N}$, $x_i \in \mathcal{X}$ give rise to a positive definite Gram matrix is called a positive definite kernel. Similarly, a function $k : \mathcal{X}^2 \to \mathbb{R}$ which for all $n \in \mathbb{N}$ and distinct $x_i \in \mathcal{X}$ gives rise to a strictly positive definite Gram matrix is called a strictly positive definite kernel.*

The RKHS is better explained in the following way. Define a map from $\mathcal{X}$ into the space of functions mapping $\mathcal{X}$ to $\mathcal{H}(k)$, via

$$\begin{aligned} \phi :& \mathcal{X} \to \mathcal{H}(k) \\ & x \to k(x, \cdot) \end{aligned}$$

Here, $\phi_x(\cdot) = k(x, \cdot)$ denotes the function that assigns the value $k(x, x')$ to $x' \in \mathcal{X}$.

We next construct a dot product space containing the images of the inputs under ϕ. To this end, considering the kernel function $k(x, x')$, suppose for n points, we fix one of the variables to have $k(x_1, x'), \ldots, k(x_n, x')$. There are all functions of the variable x'.

$$\mathcal{H}(k) = \{f(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)\} \tag{24}$$

Here, $n \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$ and $x_i \in \mathcal{X}$ are arbitrary. RKHS is a function space which is the set of all possible linear combination of these functions (Kimeldorf and Wahba, 1971). This equation shows that the bases of an RKHS are kernels, hence every function in the RKHS can be written as a linear combination.

Consider two functions in this space represented as $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$, and $g = \sum_{j=1}^n \beta_j k(x_j, \cdot)$, the inner product in RKHS is defined as

$$\begin{aligned}\langle f, g\rangle_{\mathcal{H}(k)} &= \left\langle \sum_{i=1}^n \alpha_i k(x_i, \cdot), \sum_{j=1}^n \beta_j k(x_j, \cdot) \right\rangle_{\mathcal{H}(k)} \\ &= \sum_{i,j=1}^n \alpha_i \beta_j k(x_i, x_j)\end{aligned}$$

The feature map $\phi_x(\cdot)$ is a (possibly infinite-dimensional) vector whose elements are

$$\phi_x(\cdot) = \left(\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots, \right)^\top = \Phi(x)$$

where $\{\lambda_j; j = 1, 2, \ldots\}$ and $\{\psi_j; j = 1, 2, \ldots\}$ are eigenvalues and eigenfunctions of the following eigen problem:

$$\int k(x, x')\psi_j(x')dx' = \lambda_j \psi(x)$$

See Minh et al. (2006) for details. One can understand the kernel as a similarity measurement, since the kernel can be expressed as an inner product, which is a measure of similarity in terms of angles of vectors:

$$\begin{aligned}k(x, x') &= \langle \phi_x(\cdot), \phi_{x'}(\cdot)\rangle_{\mathcal{H}(k)} \\ &= \Phi(x)^\top \Phi(x')\end{aligned}$$

Hence, the relative similarity of inputs is known by the kernel. However, in most of kernels, we cannot find an explicit expression for the feature map. Therefore, the exact location of inputs to RKHS is not necessarily known but the relative similarity, which is the kernel, is known.

B. Auxiliary Lemmas

Lemma 18 *Suppose Assumption A5 holds, we have*

$$\left\| \frac{1}{n}\sum_{s=1}^n g(z_s; \hat{\theta})k(x_s, x_j) - \mathbb{E}_{XZ}\left(g(Z : \theta_0)k(X, x_j)\right) \right\| = O_p(1/\sqrt{n}) + O_p(\|\hat{\theta} - \theta_0\|)$$

Proof

$$\begin{aligned}\frac{1}{n}\sum_{s=1}^n g(z_s; \hat{\theta})k(x_s, x_j) &= \frac{1}{n}\sum_{s=1}^n g(z_s; \theta_0)k(x_s, x_j) + \frac{1}{n}\sum_{s=1}^n G(z_s; \bar{\theta})k(x_s, x_j)(\hat{\theta} - \theta_0) \\ &= I_{1,n} + I_{2,n}(\hat{\theta} - \theta_0)\end{aligned}$$

Observe that

$$\|I_{1,n} - \mathbb{E}g(Z;\theta_0)k(X,x_j)\| = O_p(1/\sqrt{n})$$
$$\|I_{2,n} - \mathbb{E}G(Z;\theta_0)k(X,x_j)\| = O_p(1/\sqrt{n})$$

and

$$\begin{aligned} I_{2,n}(\hat{\theta}-\theta_0) &= \mathbb{E}G(Z;\theta_0)k(X,x_j)O_p(\|\hat{\theta}-\theta_0\|) + O_p(1/\sqrt{n}\|\hat{\theta}-\theta_0\|) \\ &= O_p(\|\hat{\theta}-\theta_0\|) \end{aligned}$$

■

Lemma 19 *Suppose Assumption 5 holds, we have*

$$\begin{aligned} \left\| \frac{1}{n(n-1)} \sum_{s\neq k} g(z_s;\hat{\theta})k(x_s,x_k)g^\top(z_k;\hat{\theta}) - \mathbb{E}\left(g(Z;\theta)k(X,X')g^\top(Z',\theta_0)\right) \right\| =& O_p(1/\sqrt{n}) \\ &+ O_p(\|\hat{\theta}-\theta_0\|) \end{aligned}$$

Proof

$$\begin{aligned} \frac{1}{n(n-1)} \sum_{s\neq k} g(z_s;\hat{\theta})k(x_s,x_k)g^\top(z_k;\hat{\theta}) =& \frac{1}{n(n-1)} \sum_{s\neq k} g(z_s;\theta_0)k(x_s,x_k)g^\top(z_k;\theta_0) \\ &+ \frac{1}{n(n-1)} \sum_{s\neq k} g(z_s;\theta_0)k(x_s,x_k)\left(\nabla_\theta g(z_k;\bar{\theta})(\hat{\theta}-\theta_0)\right)^\top \\ &+ \frac{1}{n(n-1)} \sum_{s\neq k} g(z_k;\theta_0)k(x_s,x_k)\left(\nabla_\theta g(z_s;\bar{\theta})(\hat{\theta}-\theta_0)\right)^\top \\ &+ \frac{1}{n(n-1)} \sum_{s\neq k} \nabla_\theta g(z_s;\bar{\theta})(\hat{\theta}-\theta_0)k(x_s,x_k)\left(\nabla_\theta g(z_k;\bar{\theta})(\hat{\theta}-\theta_0)\right)^\top \\ &= I_{1,n} + I_{21,n} + I_{22,n} + I_{3,n} \end{aligned}$$

It is clear that

$$I_{1,n} = \mathbb{E}\left(g(Z;\theta)k(X,X')g^\top(Z',\theta_0)\right) + O_p(1/\sqrt{n})$$

$$I_{21,n} = I_{22,n} = \mathbb{E}\left(\nabla_\theta g(Z;\theta_0)(\hat{\theta}-\theta_0)k(X,X')g^\top(Z';\theta_0)\right) + O_p(1/\sqrt{n}) = O_p(\|\hat{\theta}-\theta_0\|) + O_p(1/\sqrt{n})$$

$$\begin{aligned} I_{3,n} &= \mathbb{E}\left(\nabla_\theta g(Z;\theta_0)(\hat{\theta}-\theta_0)k(X,X')\left(\nabla_\theta g(Z';\theta_0)(\hat{\theta}-\theta_0)\right)^\top\right) + O_p(1/\sqrt{n}) \\ &= O_p(\|\hat{\theta}-\theta_0\|^2) + O_p(1/\sqrt{n}) \end{aligned}$$

Putting all pieces together, we yield what is asserted. ■

Lemma 20 *Suppose Assumption 5 holds, we have*

$$\|\Gamma_{n,\hat{\theta}}^{-1} - \Gamma_{\theta_0}^{-1}\| = O_p(1/\sqrt{n}) + O_p(\|\hat{\theta} - \theta_0\|)$$

Proof Observe that

$$\begin{aligned}
\Gamma_{n,\hat{\theta}} - \Gamma_{\theta_0} &= \frac{1}{n}\sum_{s=1}^{n} g(z_s;\theta_0)g^\top(z_s;\theta_0) - \mathbb{E}g(Z;\theta_0)g^\top(Z;\theta_0) \\
&= \frac{2}{n}\sum_{s=1}^{n}(g(z_s;\hat{\theta}) - g(z_s;\theta_0))g^\top(z_s;\theta_0) \\
&= \frac{1}{n}\sum_{s=1}^{n}(g(z_s;\hat{\theta}) - g(z_s;\theta_0))(g(z_s;\hat{\theta}) - g(z_s;\theta_0))^\top \\
&= O_P(1/\sqrt{n}) + O_p(\|\hat{\theta} - \theta_0\|) + O_p(\|\hat{\theta} - \theta_0\|^2) \\
&= O_P(1/\sqrt{n}) + O_p(\|\hat{\theta} - \theta_0\|)
\end{aligned}$$

By the continuous mapping theorem, the above fact yields:

$$\Gamma_{n,\hat{\theta}}^{-1} - \Gamma_{\theta_0}^{-1} = o_p(1)$$

Furthermore, we have the decomposition:

$$\Gamma_{n,\hat{\theta}}^{-1} - \Gamma_{\theta_0}^{-1} = -\Gamma_{n,\hat{\theta}}^{-1}(\Gamma_{n,\hat{\theta}} - \Gamma_{\theta_0})\Gamma_{\theta_0}^{-1}$$

Putting everything together, we have the desired result. ■

C. Verifications and Proofs

Verify the Riesz Representor of (5)

Proof Note that

$$\begin{aligned}
\mathcal{C}_\theta h &= \int \varepsilon(z;\theta)h(x)dP_{XZ}(x,z) \\
&= \int \varepsilon(z;\theta)\left\langle h, \phi_x(\cdot)\right\rangle_{\mathcal{H}(k)} dP_{XZ}(x,z) \\
&= \int \left\langle h, \varepsilon(z;\theta)\phi_x(\cdot)\right\rangle_{\mathcal{H}(k)} dP_{XZ}(x,z) \\
&= \left\langle h, \mathbb{E}_{XZ}\left(\varepsilon(Z;\theta)\phi_X(\cdot)\right)\right\rangle_{\mathcal{H}(k)} \\
&= \left\langle h, \boldsymbol{\mu}_\theta\right\rangle_{\mathcal{H}(k)}
\end{aligned}$$

To use Riesz's theorem, we need to show that the operator is bounded. Let $\|\mathcal{C}_\theta\|$ be the operator norm,

$$\|\mathcal{C}_\theta\| = \sup_{\|h\|_{\mathcal{H}(k)}\leq 1} C_\theta h = \sup_{\|h\|_{\mathcal{H}(k)}\leq 1} \langle h, \boldsymbol{\mu}_\theta\rangle_{\mathcal{H}(k)} = \left\langle \frac{\boldsymbol{\mu}_\theta}{\|\boldsymbol{\mu}_\theta\|_{\mathcal{H}(k)}}, \boldsymbol{\mu}_\theta \right\rangle_{\mathcal{H}(k)} = \|\boldsymbol{\mu}_\theta\|_{\mathcal{H}(k)}$$

with

$$\|\boldsymbol{\mu}_\theta\|^2_{\mathcal{H}(k)} = \mathbb{E}\left(\varepsilon(Z;\theta)k(X,X')\varepsilon(Z';\theta)\right) < \infty$$

by Assumption (A4). Here, (X', Z') is an independent copy of (X, Z). Furthermore, by Assumption (A2),

$$|\mathcal{C}_\theta h| \leq \|h\|_{\mathcal{H}(k)}\|\mathcal{C}_\theta\|_{\mathcal{H}(k)} < \infty$$

Thus, the $\mathcal{C}_\theta$ is a bounded linear operator. By Riesz's representation theorem, $\boldsymbol{\mu}_\theta$ is the unique representer of $\mathcal{C}_\theta$ in $\mathcal{H}(k)$. ■

Verify the Reproducing Property of (11)

Proof

$$\begin{aligned}
\left\langle \mathcal{P}\phi_x(\cdot), \phi'_x(\cdot)\right\rangle_{\mathcal{H}(k)} &= k(x,x') - \left\langle g^\top(z;\theta_0)\Gamma^{-1}_{\theta_0}\mathbb{E}_{XZ}\left(g(Z;\theta_0)\phi_X(\cdot)\right), \phi'_x(\cdot)\right\rangle_{\mathcal{H}(k)} \\
&= k(x,x') - g^\top(z;\theta_0)\Gamma^{-1}_{\theta_0}\mathbb{E}_{XZ}\left(g(Z;\theta_0)\left\langle\phi_X(\cdot), \phi'_x(\cdot)\right\rangle_{\mathcal{H}(k)}\right) \\
&= k(x,x') - g^\top(z;\theta_0)\Gamma^{-1}_{\theta_0}\mathbb{E}_{XZ}\left(g(Z;\theta_0)k(X,x')\right)
\end{aligned}$$

■

Proof of Lemma (6)

Proof

$$\begin{aligned}
\hat{k}_p(x_i,x_j) = k(x_i,x_j) &- g^\top(z_i;\hat{\theta})\Gamma^{-1}_{n,\hat{\theta}}\left(\frac{1}{n}\sum_{s=1}^n g(z_s;\hat{\theta})k(x_s,x_j)\right) \\
&- g^\top(z_j;\hat{\theta})\Gamma^{-1}_{n,\hat{\theta}}\left(\frac{1}{n}\sum_{s=1}^n g(z_s;\hat{\theta})k(x_s,x_i)\right) \\
&+ g^\top(z_i;\hat{\theta})\Gamma^{-1}_{n,\hat{\theta}}\left(\frac{1}{n(n-1)}\sum_{s\neq k}^n g(z_s;\hat{\theta})k(x_s,x_k)g^\top(z_k;\hat{\theta})\right)\Gamma^{-1}_{n,\hat{\theta}}g(z_j;\hat{\theta})
\end{aligned}$$

where

$$\Gamma_{n,\hat{\theta}} = \frac{1}{n}\sum_{s=1}^{n} g(z_s;\hat{\theta})g^\top(z_s;\hat{\theta})$$

Hence,

$$\hat{k}_p(x_i,x_j) = k(x_i,x_j) - g^\top(z_i;\hat{\theta})\Gamma^{-1}_{n,\hat{\theta}}I_{11,n} - g^\top(z_j;\hat{\theta})\Gamma^{-1}_{n,\hat{\theta}}I_{12,n} + g^\top(z_i;\hat{\theta})\Gamma^{-1}_{n,\hat{\theta}}I_{2,n}\Gamma^{-1}_{n,\hat{\theta}}g(z_j;\hat{\theta})$$

By Lemmas 18, 19 and 20, we have

- $g^\top(z_i;\hat{\theta}) = g(z_i;\theta_0) + O_p(\|\hat{\theta}-\theta_0\|)$
- $\Gamma^{-1}_{n,\hat{\theta}} = \Gamma^{-1}_{\theta_0} + O_p(1/\sqrt{n}) + O_p(\|\hat{\theta}-\theta_0\|)$
- $I_{11,n} = \mathbb{E}g(Z;\theta_0)k(X,x_j) + O_p(1/\sqrt{n}) + O_p(\|\hat{\theta}-\theta_0\|)$
- $I_{12,n} = \mathbb{E}g(Z;\theta_0)k(X,x_i) + O_p(1/\sqrt{n}) + O_p(\|\hat{\theta}-\theta_0\|)$
- $I_{2,n} = \mathbb{E}\left(g(Z;\theta)k(X,X')g^\top(Z',\theta_0)\right) + O_p(1/\sqrt{n}) + O_p(\|\hat{\theta}-\theta_0\|)$

Putting all pieces together, we have

$$\hat{k}_p(x_i,x_j) = k_p(x_i,x_j) + O_p(1/\sqrt{n}) + O_p(\|\hat{\theta}-\theta_0\|)$$

Now, we are ready to show the expanding result.

$$\begin{aligned}
n\widehat{\mathbb{M}}^2_p(\hat{\theta}) &= \frac{n}{n(n-1)}\sum_{i\neq j}\varepsilon(z_i;\theta_0)k_p(x_i,x_j)\varepsilon(z_j,\theta_0) \\
&\quad + \frac{2n}{n(n-1)}\sum_{i\neq j}\varepsilon(z_i;\theta_0)k_p(x_i,x_j)g^\top(z_j;\bar{\theta})(\hat{\theta}-\theta_0) \\
&\quad + \sqrt{n}(\hat{\theta}-\theta_0)^\top\frac{1}{n(n-1)}\sum_{i\neq j}g(z_i;\bar{\theta})k_p(x_i,x_j)g^\top(z_j;\bar{\theta})\sqrt{n}(\hat{\theta}-\theta_0) + O_p(1/\sqrt{n}) \\
&= nA^{(p)}_{1,n} + 2nA^{(p)}_{2,n}(\hat{\theta}-\theta_0) + \sqrt{n}(\hat{\theta}-\theta_0)^\top A^{(p)}_{3,n}\sqrt{n}(\hat{\theta}-\theta_0) \\
&\quad + O_p(1/\sqrt{n}) + O_p(\|\hat{\theta}-\theta_0\|)
\end{aligned}$$

where $\bar{\theta} = \gamma\hat{\theta} + (1-\gamma)\theta_0, \gamma \in (0,1)$, and the lasts term $(O_p(1/\sqrt{n}) + O_p(\|\hat{\theta}-\theta_0\|))$ comes from the fact that $1/(n(n-1))\sum_{i\neq j}\varepsilon(z_i;\theta_0)\varepsilon(z_j,\theta_0)$ is a degenerate U-statistic, hence,

$$(O_p(1/\sqrt{n}) + O_p(\|\hat{\theta}-\theta_0\|))\frac{1}{n-1}\sum_{i\neq j}\varepsilon(z_i;\theta_0)\varepsilon(z_j,\theta_0) = O_p(1/\sqrt{n}) + O_p(\|\hat{\theta}-\theta_0\|)$$

One can easily check that $A^{(p)}_{1,n}, A^{(p)}_{2,n}$ and $A^{(p)}_{3,n}$ are degenerate U-statistic, and hence $A^{(p)}_{1,n}, A^{(p)}_{2,n}, A^{(p)}_{3,n} = O_P(1/n)$. Thus, we have

$$\begin{aligned}
n\widehat{\mathbb{M}}^2_p(\hat{\theta}) =& \frac{n}{n(n-1)}\sum_{i\neq j}\varepsilon(z_i;\theta_0)k_p(x_i,x_j)\varepsilon(z_j,\theta_0) + O_p(\|\hat{\theta}-\theta_0\|) + O_p(\|\hat{\theta}-\theta_0\|^2) \\
&+ O_p(1/\sqrt{n})
\end{aligned}$$

■

Proof of Theorem 5.

Proof Suppose that $\mu^{(p)}_{\theta_1} = \mu^{(p)}_{\theta_2}$ and let $\delta(x) = \mathcal{E}(x;\theta_1) - \mathcal{E}(x,\theta_2)$. Then we have

$$\begin{aligned}
\|\mu^{(p)}_{\theta_1} - \mu^{(p)}_{\theta_2}\|^2_{\mathcal{H}(k_p)} &= \left\| \int \xi^{(p)}_{\theta_1}(x,z)dP_{XZ}(x,z) - \int \xi^{(p)}_{\theta_2}(x,z)dP_{XZ}(x,z) \right\|^2_{\mathcal{H}(k_p)} \\
&= \left\| \int \mathcal{E}(x;\theta_1)\phi^{(p)}_x(\cdot)dP_X(x) - \int \mathcal{E}(x;\theta_2)\phi^{(p)}_x(\cdot)dP_X(x) \right\|^2_{\mathcal{H}(k_p)} \\
&= \left\| \int (\mathcal{E}(x;\theta_1) - \mathcal{E}(x;\theta_2))\phi^{(p)}_x(\cdot)dP_X(x) \right\|^2_{\mathcal{H}(k_p)} \\
&= \int\int \delta(x)k_p(x,x')\delta(x')dP_X(x)dP_{X'}(x') = 0
\end{aligned}$$

where X' is an independent copy of X. Since $k_p(\cdot,\cdot)$ is ISPD kernel and the assumption that $\delta(x)$ is not colinear with $g(Z;\theta_0)$, it follows that the function $\varphi(x) = \delta(x)p_X(x)$ has zero L2-norm, i.e., $\|\varphi\|_2^2 = 0$ where p_X denotes the density of P_X. As a result, $\delta(x) = 0$ a.s P_X implying that $P_X(B_0) = 1$ where $B_0 = \{x \in \mathcal{X} : \mathcal{E}(x;\theta_1) - \mathcal{E}(x;\theta_2) = 0\}$. Therefore, $\mathcal{E}(x;\theta_1) = \mathcal{E}(x;\theta_2)$ for P_X almost surely. ■

Proof of Theorem 9

Proof

Under local alternatives, $\varepsilon(z;\hat\theta) = \varepsilon(z;\theta_0) + (\hat\theta - \theta_0)^\top g(z;\bar\theta) + R(x)/\sqrt{n}$. Hence,

$$\begin{aligned}
n\widehat{\mathbb{M}}^2_n(\hat\theta) &= \frac{n}{n(n-1)} \sum_{1\le i\ne j\le n} \varepsilon(z_i;\theta_0)\hat k_p(x_i,x_j)\varepsilon(z_j;\theta_0) \\
&+ \frac{2n}{n(n-1)} \sum_{1\le i\ne j\le n} \varepsilon(z_i;\theta_0)\hat k_p(x_i,x_j)g^\top(z_j;\bar\theta)(\hat\theta - \theta_0) \\
&+ \sqrt{n}(\hat\theta-\theta_0)^\top \left(\frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} g(z_i;\bar\theta)\hat k_p(x_i,x_j)g^\top(z_j;\bar\theta) \right) \sqrt{n}(\hat\theta-\theta_0) \\
&+ \frac{2n}{n(n-1)} \sum_{1\le i\ne j\le n} \varepsilon(z_i;\theta_0)\hat k_p(x_i,x_j)\frac{R(x_j)}{\sqrt{n}} \\
&+ \frac{2n}{n(n-1)} \sum_{1\le i\ne j\le n} (\hat\theta-\theta_0)^\top g(z_i;\bar\theta)\hat k_p(x_i,x_j)\frac{R(x_j)}{\sqrt{n}} \\
&+ \frac{n}{n(n-1)} \sum_{1\le i\ne j\le n} \frac{R(x_i)}{\sqrt{n}}\hat k_p(x_i,x_j)\frac{R(x_j)}{\sqrt{n}} \\
&= nA_{1,n} + 2nA_{2,n}(\hat\theta-\theta_0) + \sqrt{n}(\hat\theta-\theta_0)^\top A_{3,n}\sqrt{n}(\hat\theta-\theta_0) + 2\sqrt{n}A_{4,n} \\
&+ 2\sqrt{n}(\hat\theta-\theta_0)^\top A_{5,n} + A_{6,n}
\end{aligned}$$

Note that

$$nA_{1,n} \xrightarrow{d} \sum_{k=1}^{\infty} \tau_k^{(p)}(W_k^2 - 1)$$

by Theorem 7. $A_{2,n}$ is a degenerate U-statistic by the orthogonality argument, hence $2nA_{2,n}(\hat{\theta} - \theta_0) = O_p(\|\hat{\theta} - \theta_0\|) = o_p(1)$.

Furthermore,

$$A_{3,n} = O_p(1/n)$$

,

$$A_{l,n} = O_p(1/\sqrt{n}), \qquad l = 4, 5$$

by the orthogonality between $k_p(\cdot, \cdot)$ and $g(z; \theta)$ and the fact that $\hat{k}_p(x, x') = k_p(x, x') + O_p(1/\sqrt{n}) + O_p(\|\hat{\theta} - \theta_0\|)$. Thus, by Serfling (1980)

$$\begin{aligned}
&\sqrt{n}(\hat{\theta} - \theta_0)^\top A_{3,n}\sqrt{n}(\hat{\theta} - \theta_0) + 2\sqrt{n}A_{4,n} + 2\sqrt{n}(\hat{\theta} - \theta_0)^\top A_{5,n} \\
&\xrightarrow{d} 2N\left(0, 4\mathrm{Var}_{X,Z}\left(\mathbb{E}_{X',Z'}\varepsilon(Z; \theta_0)k_p(X, X')R(X')\right)\right)
\end{aligned}$$

Lastly,

$$A_{6,n} = \mathbb{E}\left(R(X)k_p(X, X')R(X')\right) + O_p(1/\sqrt{n})$$

Putting these pieces together, we yield what is asserted. ■

Proof of Lemma 10

Proof

$$\begin{aligned}
\mathbb{M}_p^2(\theta_0) &= \mathbb{E}\left(\varepsilon(Z; \theta_0)k_p(X, X')\varepsilon(Z'; \theta_0) + \frac{R(X)R(X')}{n}k_p(X, X') + 2\varepsilon(Z; \theta_0)\frac{R(X')}{\sqrt{n}}k_p(X, X')\right) \\
&= \sum_{j\geq 1} \lambda_j \left(\mathbb{E}\varepsilon(Z; \theta_0)e_j(X)\right)^2 + \sum_{j\geq 1} \lambda_j \left(\mathbb{E}\frac{R(X)}{\sqrt{n}}e_j(X)\right)^2 \\
&\quad + 2\sum_{j\geq 1} \lambda_j \mathbb{E}\varepsilon(Z; \theta_0)e_j(X)\mathbb{E}\left(\frac{R(X)}{\sqrt{n}}e_j(X)\right)
\end{aligned}$$

Recall $R(x) = \sum_{s\geq 1} \alpha_s e_s(x)$, we can further conclude that

$$\begin{aligned}
\mathbb{M}_p^2(\theta_0) &= \sum_{j\geq 1} \lambda_j \left(\mathbb{E}\varepsilon(Z; \theta_0)e_j(X)\right)^2 + \lambda_j \frac{\alpha_j^2}{n}\mathbb{E}e_j^2(X) + 2\lambda_j \mathbb{E}\varepsilon(Z; \theta_0)e_j(X)\frac{\alpha_j}{\sqrt{n}}\mathbb{E}e_j^2(X) \\
&= \mathbb{E}\left(\varepsilon(Z; \theta_0)k_p(X, X')\varepsilon(Z'; \theta_0)\right) + \lambda_j \frac{\alpha_j^2}{n} + 2\lambda_j \mathbb{E}\left(\varepsilon(Z; \theta_0)e_j(X)\right)\frac{\alpha_j}{\sqrt{n}}
\end{aligned}$$

■

Proof of Theorem 11.

Proof To stream line the presentation, let

$$\hat{f}_\theta(u_i, u_j) = \varepsilon(z_i;\theta)v_i\hat{k}_p(x_i, x_j)\varepsilon(z_j;\theta)v_j$$

$$f_\theta(u_i, u_j) = \varepsilon(z_i;\theta)v_i k_p(x_i, x_j)\varepsilon(z_j;\theta)v_j$$

where $u_i = (v_i, x_i, z_i)$. Furthermore, let $(x, z)^{(n)} = \{(x_i, z_i); i = 1, \ldots, n\}$, $\hat{\theta} - \theta^* = O_p(1/\sqrt{n})$ under different hypotheses, and

$$k_p(x, x') = k(x, x') - g^\top(z;\theta^*)\Gamma_{\theta^*}^{-1}\mathbb{E}_{(X,Z)}\left(g(Z;\theta^*)k(X, x')\right)$$

By the fact that $\hat{k}_p(\cdot,\cdot) = k_p(\cdot,\cdot) + O_p(\|\hat{\theta} - \theta^*\|) + O_p(1/\sqrt{n})$, we have

$$\begin{aligned}
n\widehat{\mathbb{M}}_p^{2,*}(\hat{\theta}) &= \frac{n}{n(n-1)}\sum_{i\neq j}\hat{f}_{\hat{\theta}}(u_i, u_j) \\
&= \frac{n}{n(n-1)}\sum_{i\neq j} f_{\theta_0}(u_i, u_j) + o_p(1) \\
&= n\mathbb{M}_p^{2,*}(\theta_0) + o_p(1)
\end{aligned}$$

Let

$$T_n = \frac{1}{n}\sum_{i\neq j} f_{\hat{\theta}}(u_i, u_j)$$

we have $n\mathbb{M}_p^{2,*}(\theta_0) = \frac{n}{n-1}T_n$, the goal is to show that

$$T_n \xrightarrow{d,*} Y = \sum_{k=1}^{\infty}\tau_k^{(p)}(W_k^2 - 1)$$

We shall carry this out by the method of characteristic functions, i.e., to show that

$$\mathbb{E}(e^{i\omega T_n}|(x, z)^{(n)}) \to \mathbb{E}(e^{i\omega Y}), \quad n \to \infty, \forall\omega$$

Denote $\{\rho_k(\cdot)\}$ as the orthonormal eigenfunctions corresponding to the eigenvalues $\{\tau_k^{(p)}\}$ defined in connection with $\varepsilon(z;\theta_0)k_p(x, x')\varepsilon(z';\theta_0)$. Thus,

$$f_{\theta_0}(u_1, u_2) = \sum_{k\geq 1}\tau_k^{(p)}v_1v_2\rho_k(y_1)\rho_k(y_2)$$

with $y_1 = (x_1, z_1)$

Thus, T_n might be expressed as

$$T_n = \frac{1}{n}\sum_{i\neq j}\sum_{k\geq 1}\tau_k^{(p)}v_iv_j\rho_k(y_i)\rho_k(y_j)$$

Now put

$$T_{nK} = \frac{1}{n} \sum_{i \neq j} \sum_{k=1}^{K} \tau_k^{(p)} v_i v_j \rho_k(y_i) \rho_k(y_j)$$

Using the equality $|e^{iz} - 1| < |z|$, we have for an arbitrary $\delta > 0$, there exists a K such that

$$\begin{aligned}
|\mathbb{E}(e^{i\omega T_n}|(x,z)^{(n)}) - \mathbb{E}(e^{i\omega T_{nK}}|(x,z)^{(n)})| &\leq \mathbb{E}((|e^{i\omega T_n} - e^{i\omega T_{nK}}|)|(x,z)^{(n)}) \\
&\leq |\omega| \mathbb{E}((|T_n - T_{nK}|)|(x,z)^{(n)}) \\
&\leq |\omega| \left(\mathbb{E}(T_n - T_{nK})^2 |(x,z)^{(n)} \right)^{1/2}
\end{aligned}$$

Observe that $T_n - T_{nK}$ is of the form of a U-statistic, that is,

$$T_n - T_{nK} = \frac{2}{n} \binom{n}{2} U_{nK}$$

where

$$U_{nK} = \binom{n}{2}^{-1} \sum_{i \neq j} g_K(u_i, u_j)$$

with

$$g_K(u_1, u_2) = \sum_{k=K+1}^{\infty} \tau_k^{(p)} \rho_k(y_1) \rho_k(y_2) v_1 v_2$$

Note that

$$\begin{aligned}
\mathbb{E}\left(U_{nK}^2 |(x,z)^{(n)} \right) &= \left[\sum_{k=K+1}^{\infty} \left(\binom{n}{2}^{-1} \sum_{i \neq j} \tau_k^{(p)} \rho_k(y_i) \rho_k(y_j) \right) \right]^2 \\
&= \left(\sum_{k=K+1}^{\infty} U_{nk}^* \right)^2
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left((T_n - T_{nK})^2 |(x,z)^{(n)} \right) &= (n-1)^2 \binom{n}{2}^{-1} \left(\sum_{k=K+1}^{\infty} U_{nk}^* \right)^2 \\
&\leq 2 \left(\sum_{k=K+1}^{\infty} U_{nk}^* \right)^2
\end{aligned}$$

Since

$$\left(\sum_{k=1}^{\infty} U_{nk}^* \right)^2 = \left(\binom{n}{2}^{-1} \sum_{i \neq j} \varepsilon(z_i; \theta^*) k_p(x_i, x_j) \varepsilon(z_j; \theta^*) \right)^2 < \infty$$

One can fix ω and let $\delta > 0$ be given, then choose and fix K large enough that

$$|\omega| \left(2 \left(\sum_{k=K+1}^{\infty} U_{nk}^* \right)^2 \right)^{1/2} < \delta$$

Thus we have

$$|\mathbb{E}(e^{i\omega T_n}|(x,z)^{(n)}) - \mathbb{E}(e^{i\omega T_{nK}}|(x,z)^{(n)})| < \delta \tag{25}$$

Next we show that $T_{nK}|(x,z)^{(n)} \xrightarrow{d} Y_k = \sum_{k=1}^{K} \tau_k^{(p)}(W_k^2 - 1)$. Let

$$W_{kn} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i \rho_k(y_i); \quad Z_{kn} = \frac{1}{n} \sum_{i=1}^{n} v_i^2 \rho_k^2(y_i)$$

then,

$$T_{nK} = \sum_{k=1}^{K} \tau_k^{(p)}(W_{nk}^2 - Z_{nk})$$

Notice that

$$\mathbb{E}(W_{nk}|(x,z)^{(n)}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \rho_k(y_i)\,(\mathbb{E}V) = 0$$

and

$$Cov(W_{jn}, W_{kn}|(x,z)^{(n)}) = \frac{1}{n} \sum_{i=1}^{n} \rho_k(y_i)\rho_j(y_i) \xrightarrow{p} \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}$$

Therefore, by Lindeberg-Levy CTL,

$$(W_{1n}, \ldots, W_{Kn})\,|(x,z)^{(n)} \xrightarrow{d} N(0, I_K)$$

Furthermore, by SLLN,

$$(Z_{1n}, \ldots, Z_{Kn})\,|(x,z)^{(n)} \xrightarrow{p} (1, \ldots, 1)$$

Thus,

$$T_{nK}|(x,z)^{(n)} \xrightarrow{d} Y_k = \sum_{k=1}^{K} \tau_k^{(p)}(W_k^2 - 1)$$

and for all n sufficiently large

$$|\mathbb{E}(e^{i\omega T_{nK}}|(x,z)^{(n)}) - \mathbb{E}e^{i\omega Y_K}| < \delta \tag{26}$$

Finally, denote Y as the limit in mean square of Y_k as $K \to \infty$. Then

$$\begin{aligned} |\mathbb{E}e^{i\omega Y_K} - \mathbb{E}e^{i\omega Y}| &\leq |\omega| \left[\mathbb{E}(Y - Y_K)^2\right]^{1/2} \\ &\leq |\omega| \left[\mathbb{E}(W_1^2 - 1)^2\right]^{1/2} \left[\sum_{k=K+1}^{\infty} (\tau_k^{(p)})^2\right]^{1/2} \\ &< \delta \left[\mathbb{E}(W_1^2 - 1)^2\right]^{1/2} \end{aligned} \tag{27}$$

Combining inequality equations (25),(26) and (27), we have, for any ω and any $\delta > 0$, and for all n sufficiently large,

$$|\mathbb{E}(e^{i\omega T_n}|(x,z)^{(n)}) - \mathbb{E}(e^{i\omega Y})| \leq \delta \left(2 + \left[\mathbb{E}(W_1^2 - 1)^2\right]^{1/2}\right)$$

Thus,

$$T_n \xrightarrow{d,*} Y$$

■

Proof of Theorem 12.

Proof Essentially, we need to show the uniformly convergence of $\widehat{R}_V(\theta)$, the rest of the consistency proof follows immediately from Theorem 2.1 of Newey and McFadden (1994).

To prove that $\sup_{\theta\in\Theta}|\widehat{R}_V(\theta) - R_k(\theta)| \overset{p}{\to} 0$, we need to show that (1) $f_\theta(v, v')$ is continuous at each θ with probability one; and (2) $\mathbb{E}_{V,V'}\left(\sup_{\theta\in\Theta}|f_\theta(V, V')|\right) < \infty$, and $\mathbb{E}_{V,V}\left(\sup_{\theta\in\Theta}|f_\theta(V, V)|\right) < \infty$ (Lemma 8.5 of Newey and McFadden (1994)).

To this end, we can check that

$$\begin{aligned}
|f_\theta(v, v')| &= |\varepsilon(z;\theta)k(x, x')\varepsilon(z';\theta)| \\
&\le |\varepsilon(z;\theta)||\varepsilon(z';\theta)||k(x, x')| \\
&\le |\varepsilon(z;\theta)||\varepsilon(z';\theta)|\sqrt{k(x, x)k(x', x')}
\end{aligned}$$

Since Θ is compact, $\mathbb{E}\left(|Y|\right) < \infty$, and $\mathbb{E}_\theta\left(Y|X\right) < \infty$, we have $|\varepsilon(z;\theta)| < \infty$ for all $\theta \in \Theta$. Furthermore, $k(\cdot,\cdot)$ is bounded by Assumption A4, we have $f_\theta(v, v') < \infty$ and thus $f_\theta(v, v')$ is continuous at each θ.

Next, observe that

$$\begin{aligned}
\mathbb{E}_{V,V'}\left(\sup_{\theta\in\Theta}|f_\theta(V, V')|\right) &\le \mathbb{E}\left(\sup_{\theta\in\Theta}|\varepsilon(Z;\theta)||\varepsilon(Z';\theta)|\sqrt{k(X, X)k(X', X')}\right) \\
&\le \mathbb{E}\left(\sup_{\theta\in\Theta}|\varepsilon(Z;\theta)||\varepsilon(Z';\theta)|\right)\sup_x k(x, x) \\
&= \mathbb{E}^2\left(\sup_{\theta\in\Theta}|\varepsilon(Z;\theta)|\right)\sup_x k(x, x) < \infty
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{V,V}\left(\sup_{\theta\in\Theta}|f_\theta(V, V)|\right) &\le \mathbb{E}\left(\sup_{\theta\in\Theta}(|\varepsilon(Z;\theta)|)^2\right)\sup_x k(x, x) \\
&= \mathbb{E}\left(\sup_{\theta\in\Theta}(|Y - \mathbb{E}_\theta(Y|X)|)^2\right)\sup_x k(x, x) \\
&\le 2\left(\mathbb{E}(|Y|^2) + \mathbb{E}\left(\sup_{\theta\in\Theta}|\mathbb{E}_\theta(Y|X)|^2\right)\right)\sup_x k(x, x) < \infty
\end{aligned}$$

■

Proof of Theorem 13.

Proof By Theorem 3.1 of Newey and McFadden (1994), we need to show

- (a) $\hat{\theta} - \theta_0 \xrightarrow{p} 0$;
- (b) $\widehat{R}_V(\theta)$ is twice continuously differentiable;
- (c) $\sqrt{n}\widehat{R}_V(\theta) \xrightarrow{d} N(0, 4\mathrm{Var}_V\left(\mathbb{E}^2_{V'}(\nabla_\theta f_{\theta_0}(V, V'))\right))$;
- (d) there exist $H(\theta)$ that is continuous at θ_0 and $\sup_{\theta\in\Theta}\|\nabla^2_\theta\widehat{R}_V(\theta) - \mathbb{E}\left(\nabla^2_\theta f_\theta(V, V')\right)\|_F \xrightarrow{p} 0$;
- (e) $H(\theta_0)$ is non-singular.

By consistency result and assumptions made, we only need to check conditions (c) and (d).

Since $\sqrt{n}\left(\nabla_\theta\widehat{R}_V(\theta) - \nabla_\theta\widehat{R}_U(\theta)\right) \xrightarrow{p} 0$ (see, Section 5.7.3 of Serfling (1980)), we can check the asymptotic properties of $\sqrt{n}\nabla_\theta\widehat{R}_U(\theta)$ instead. Note that

$$\nabla_\theta\widehat{R}_U(\theta) = \frac{1}{n(n-1)}\sum_{i\neq j}\nabla_\theta f_\theta(v_i, v_j)$$

$$\nabla_\theta f_\theta(v_i, v_j) = \left(g(z_i;\theta)\varepsilon(z_j;\theta) + g(z_j;\theta)\varepsilon(z_i;\theta)\right)k(x_i, x_j)$$

First, we show that

$$\sqrt{n}\nabla_\theta\widehat{R}_U(\theta) \xrightarrow{d} N(0, 4\mathrm{Var}_V\left(\mathbb{E}^2_{V'}(\nabla_\theta f_{\theta_0}(V, V'))\right))$$

The proof follows from Section 5.5.1 and 5.5.2 of Serfling (1980). We need to show (i) $\nabla_\theta\widehat{R}_U(\theta_0) \xrightarrow{p} 0$; and (ii) whether $\mathrm{Var}_V\left(\mathbb{E}^2_{V'}(f_{\theta_0}(V, V'))\right) > 0$ or not. (i) can be easily obtained by L.L.N.

To verify (ii), note that

$$\mathrm{Var}_V\left(\mathbb{E}^2_{V'}(\nabla_\theta f_{\theta_0}(V, V'))\right) = \mathbb{E}_V\left(\mathbb{E}^2_{V'}(\nabla_\theta f_{\theta_0}(V, V'))\right) \geq 0$$

where the equality hold if for any V, there is $\mathbb{E}_{V'}(\nabla_\theta f_{\theta_0}(V, V')) = 0$, i.e.,

$$\mathbb{E}_{V'}(\nabla_\theta f_{\theta_0}(V, V')) = \mathbb{E}_{V'}\left(g(Z';\theta_0)k(X, X')\right)\varepsilon(Z;\theta_0) + \mathbb{E}_{V'}\left(\varepsilon(Z';\theta_0)k(X, X')\right)g(Z;\theta_0) = 0$$

Since

$$\mathbb{E}_{V'}\left(\varepsilon(Z';\theta_0)k(X, X')\right) = \mathbb{E}_{X'}\left(\mathbb{E}(\varepsilon(Z';\theta_0)|X')k(X, X')\right) = 0$$

The equality holds if

$$\mathbb{E}_{V'}\left(g(Z';\theta_0)k(X, X')\right) = \mathbb{E}_{X'}\left(\mathbb{E}(g(Z';\theta_0)|X')k(X, X')\right) = 0$$

However, the assertion that $\mathbb{E}(g(Z';\theta_0)|X') = 0$ contradicts with the condition that H is non-singular. Thus, we conclude

$$\mathrm{Var}_V\left(\mathbb{E}^2_{V'}(\nabla_\theta f_{\theta_0}(V, V'))\right) > 0$$

Next, we show the uniform consistency of $\nabla_\theta^2 \widehat{R}_U(\theta)$. Note that

$$\nabla_\theta^2 \widehat{R}_U(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} \nabla_\theta^2 f_\theta(v_i, v_j)$$

$$\nabla_\theta^2 f_\theta(v_i, v_j) = \Big(g(z_i;\theta) g^\top(z_j;\theta) + g(z_j;\theta) g^\top(z_i;\theta) + \nabla_\theta g(z_i;\theta)\varepsilon(z_j;\theta) + \nabla_\theta g(z_j;\theta)\varepsilon(z_i;\theta) \Big) k(x_i, x_j)$$

We need to show that (i) $\nabla_\theta^2 \widehat{R}_U(\theta)$ is continuous at each θ with probability one, and (ii) there exists $\mathbb{E}\left(\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(V, V')\|_F^2\right) < \infty$ and $\mathbb{E}\left(\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(V, V)\|_F^2\right) < \infty$.

To prove (i), we exploit the triangle inequality of the Frobenius norm,

$$\begin{aligned}
\|\nabla_\theta^2 \widehat{R}_U(\theta)\|_F &\leq \Big(2\|g(z;\theta) g^\top(z';\theta)\|_F + |\varepsilon(z;\theta)| \|\nabla_\theta g(z';\theta)\|_F + |\varepsilon(z';\theta)| \|\nabla_\theta g(z;\theta)\|_F \Big) k(x, x') \\
&= d(v, v')
\end{aligned}$$

Since $\mathbb{E}_\theta(Y|X)$ is twice continuously differentiable about θ and Θ is compact, we have $\mathbb{E}_\theta(Y|X)$ bounded as well as each entry of $g(z;\theta)$ and $\nabla_\theta g(z;\theta)$ for $\|z\| < \infty$. Furthermore, since $k(\cdot,\cdot)$ is also bounded, thus, $d(v, v') < \infty$ if v, v' are bounded. We conclude then (i) must hold.

To show (ii), note that

$$\begin{aligned}
\mathbb{E}\left(\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(V, V')\|_F^2 \right) &\leq 2\mathbb{E}\left(\sup_{\theta\in\Theta} \|g(Z;\theta) g^\top(Z';\theta)\|_F + |\varepsilon(Z;\theta)| \|\nabla_\theta g(Z';\theta)\|_F \right) \sup_x k(x, x) \\
&= 2\left(\left(\mathbb{E} \sup_{\theta\in\Theta} \|g(Z;\theta)\|_F \right)^2 + \mathbb{E}\left(\sup_{\theta\in\Theta} |\varepsilon(Z;\theta)| \right) \mathbb{E}\left(\sup_{\theta\in\Theta} \|\nabla_\theta g(Z';\theta)\|_F \right) \right) \\
&\times \sup_x k(x, x) \\
&< \infty
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left(\sup_{\theta\in\Theta} \|\nabla_\theta^2 f_\theta(V, V)\|_F^2 \right) &\leq 2\mathbb{E}\left(\|g(Z;\theta) g^\top(Z;\theta)\|_F + |\varepsilon(Z;\theta)| \|\nabla_\theta g(Z;\theta)\|_F \right) \sup_x k(x, x) \\
&\leq \left(2\mathbb{E}\left(\sup_{\theta\in\Theta} \|g(Z;\theta)\|_F^2 \right) + 2\mathbb{E}\left(\sup_{\theta\in\Theta} |\varepsilon(Z;\theta)| \right) \mathbb{E}\left(\sup_{\theta\in\Theta} \|\nabla_\theta g(Z;\theta)\|_F \right) \right) \\
&\times \sup_x k(x, x) \\
&< \infty
\end{aligned}$$

Therefore, by Lemma 8.5 of Newey and McFadden (1994), we have

$$\sup_{\theta\in\Theta} \|\nabla_\theta^2 \widehat{R}_V(\theta) - \mathbb{E}\left(\nabla_\theta^2 f_\theta(V, V')\right)\|_F \xrightarrow{p} 0$$

The rest of the asymptotic normality proof follows from Theorem 3.1 of Newey and McFadden (1994). ■

Verification of Equation (23).

Proof

$$\begin{aligned}\mathbb{E}\left[\varepsilon(Z;\theta)k_p(X,X')\varepsilon(Z';\theta)\right] &= \mathbb{E}\left\langle \varepsilon(Z;\theta)\mathcal{P}\phi_X(\cdot), \varepsilon(Z';\theta)\mathcal{P}\phi_{X'}(\cdot)\right\rangle_{\mathcal{H}(k)} \\ &= \left\langle \mathbb{E}_{(X,Z)}\varepsilon(Z;\theta)\mathcal{P}\phi_X(\cdot), \mathbb{E}_{(X',Z')}\varepsilon(Z';\theta)\mathcal{P}\phi_{X'}(\cdot)\right\rangle_{\mathcal{H}(k)}\end{aligned}$$

Observe that

$$\begin{aligned}\mathbb{E}_{(X,Z)}\varepsilon(Z;\theta)\mathcal{P}\phi_X(\cdot) &= \int_{\mathcal{X},\mathcal{Z}} \varepsilon(z;\theta)\left(\phi_x(\cdot) - g^\top(z;\theta)\Gamma_{\theta_0}^{-1}\mathbb{E}_{(X,Z)}\left(g(Z;\theta)\phi_X(\cdot)\right)\right) dP_{(X,Z)}(x,z) \\ &= \int_{\mathcal{X},\mathcal{Z}} \varepsilon(z;\theta)\phi_x(\cdot) dP_{(X,Z)}(x,z) \\ &\quad - \int_{\mathcal{X},\mathcal{Z}} \varepsilon(z;\theta) g^\top(z;\theta)\Gamma_{\theta_0}^{-1}\mathbb{E}_{(X,Z)}\left(g(Z;\theta)\phi_X(\cdot)\right) dP_{(X,Z)}(x,z)\end{aligned}$$

Further analysis of the second part, we have

$$\begin{aligned}&\int_{\mathcal{X},\mathcal{Z}} \varepsilon(z;\theta) g^\top(z;\theta)\Gamma_{\theta_0}^{-1}\mathbb{E}_{(X,Z)}\left(g(Z;\theta)\phi_X(\cdot)\right) dP_{(X,Z)}(x,z) \\ &= \int_{\mathcal{X},\mathcal{Z}} \varepsilon(z;\theta) g^\top(z;\theta)\Gamma_{\theta_0}^{-1}\int_{\mathcal{X},\mathcal{Z}} g(z;\theta)\phi_x(\cdot) dP_{(X,Z)}(x,z) dP_{(X,Z)}(x,z) \\ &= \int_{\mathcal{X},\mathcal{Z}}\int_{\mathcal{X},\mathcal{Z}} \phi_x(\cdot) g^\top(z;\theta)\Gamma_{\theta_0}^{-1} g(z;\theta)\varepsilon(z;\theta) dP_{(X,Z)}(x,z) dP_{(X,Z)}(x,z) \\ &= \int_{\mathcal{X},\mathcal{Z}} \phi_x(\cdot) g^\top(z;\theta)\Gamma_{\theta_0}^{-1}\mathbb{E}_{(X,Z)}\left(g(Z;\theta)\varepsilon(Z;\theta)\right) dP_{(X,Z)}(x,z)\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}_{(X,Z)}\varepsilon(Z;\theta)\mathcal{P}\phi_X(\cdot) &= \int_{\mathcal{X},\mathcal{Z}} \varepsilon(z;\theta)\phi_x(\cdot) dP_{(X,Z)}(x,z) \\ &\quad - \int_{\mathcal{X},\mathcal{Z}} \phi_x(\cdot) g^\top(z;\theta)\Gamma_{\theta_0}^{-1}\mathbb{E}_{(X,Z)}\left(g(Z;\theta)\varepsilon(Z;\theta)\right) dP_{(X,Z)}(x,z) \\ &= \int_{\mathcal{X},\mathcal{Z}} \left(\varepsilon(z;\theta) - g^\top(z;\theta)\Gamma_{\theta_0}^{-1}\mathbb{E}_{(X,Z)}\left(g(Z;\theta)\varepsilon(Z;\theta)\right)\right)\phi_x(\cdot) dP_{(X,Z)}(x,z) \\ &= \int_{\mathcal{X},\mathcal{Z}} \varepsilon_p(z;\theta)\phi_x(\cdot) dP_{(X,Z)}(x,z) = \mathbb{E}_{(X,Z)}\varepsilon_p(Z;\theta)\phi_X(\cdot)\end{aligned}$$

Hence,

$$\mathbb{E}\left[\varepsilon(Z;\theta)k_p(X,X')\varepsilon(Z';\theta)\right] = \mathbb{E}\left[\varepsilon_p(Z;\theta)k(X,X')\varepsilon_p(Z';\theta)\right]$$

■