Kernel Conditional Moment Test via Maximum Moment Restriction

Krikamol Muandet
MPI for Intelligent Systems
Tübingen, Germany

Wittawat Jitkrittum[*]
MPI for Intelligent Systems
Tübingen, Germany

Jonas M. Kübler
MPI for Intelligent Systems
Tübingen, Germany

Abstract

We propose a new family of specification tests called kernel conditional moment (KCM) tests. Our tests are built on a novel representation of conditional moment restrictions in a reproducing kernel Hilbert space (RKHS) called conditional moment embedding (CMME). After transforming the conditional moment restrictions into a continuum of unconditional counterparts, the test statistic is defined as the maximum moment restriction (MMR) within the unit ball of the RKHS. We show that the MMR not only fully characterizes the original conditional moment restrictions, leading to consistency in both hypothesis testing and parameter estimation, but also has an analytic expression that is easy to compute as well as closed-form asymptotic distributions. Our empirical studies show that the KCM test has a promising finite-sample performance compared to existing tests.

1 INTRODUCTION

Many problems in causal inference, economics, and finance are often formulated as a conditional moment restriction (CMR): for correctly specified models, the conditional mean of certain functions of data is almost surely equal to zero [1, 38]. Rational expectation models—widely used in many fields of macroeconomics—specify how economic agents exploit available information to form their expectations in terms of conditional moments [35]. Recent advances in causal machine learning also rely on the CMR including a generalized random forest (GRF) [6], orthogonal random forest (ORF) [41], double machine learning (DML) [13], and nonparametric instrumental variable regression [7, 30] among others; see also Hartford et al. [24], Muandet et al. [34], Singh et al. [46] and references therein.

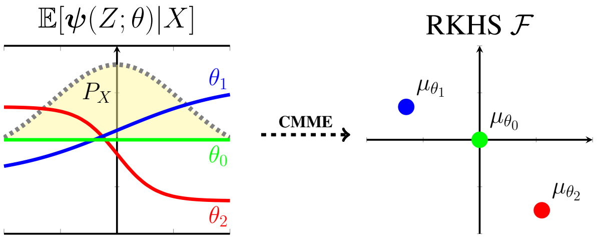


Figure 1: **Conditional moment embedding (CMME)**: The conditional moments $\mathbb{E}[\boldsymbol{\psi}(Z;\theta)|X]$ for different parameters θ are *uniquely* (P_X-almost surely) embedded into the RKHS. The RKHS norm of $\boldsymbol{\mu}_\theta$ measures to what extent these restrictions are violated and hence is used as a test statistic for conditional moment tests.

Checking the validity of these moment restrictions is the first and foremost step to ensure that a model is correctly specified which constitutes a fundamental assumption for its estimation and inference. A model misspecification often creates biases to parameter estimates, inconsistency of standard errors, and invalid asymptotic distributions that hinder our subsequent inference based on the model. An overidentifying restriction test in the generalized method of moments (GMM) framework is one of the standard approaches to test a *finite* number of *unconditional* moment conditions [22, 23]. The J-test is an example of such tests [23, 43], and numerous tests have been developed in econometrics to deal with various sources of misspecification; see, *e.g.*, Bierens [10] for a review. This paper focuses on an important class of CMR-based specification tests known as the conditional moment (CM) tests [36, 52] which have a long history in econometrics [10, 25, 55].

Testing *conditional* moment restrictions becomes more challenging as an *infinite* number of equivalent unconditional moment restrictions (UMR) must be examined simultaneously (cf. Section 3). At first, Newey [36] and

[*] Now with Google Research.

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

Tauchen [52] proposed to perform the overidentifying restriction test on a finite subset of the UMR. Unfortunately, the CM tests that rely only on a finite number of moment conditions cannot be consistent against all alternatives. Additional assumptions such as the global identification of selected moment conditions and sample-size dependent moment conditions are required to guarantee consistency [15, 18]. To overcome this limitation, Bierens [8] introduced the first consistent CM tests—known as integrated conditional moment (ICM) tests—by checking *all* moment conditions simultaneously [11]. However, the ICM test depends on parametric weighting functions and nuisance parameters that limit its practical use. An alternative class of consistent CM tests, known as smooth tests, employ nonparametric kernel estimation [31, 56] which also forms a basis for the generalized empirical likelihood approach [16, 54]. However, they have non-trivial power only against local alternatives that approach the null at a slower rate than $1/\sqrt{n}$, and are susceptible to the curse of dimensionality (cf. Section 5 for the discussion).

Inspired by a surge of kernel-based tests [14, 21, 32], we propose to embed the CMR in a reproducing kernel Hilbert space (RKHS). By transforming CMR into a continuum of UMR in RKHS, the test statistic is defined as the maximum moment restriction (MMR) within the unit ball of the RKHS (cf. Section 3). We then show that the MMR corresponds to the RKHS norm of a Hilbert space embedding of conditional moments. *Not only can the MMR capture all information about the original CMR, but it also has a closed-form expression that enables the practical ease of implementation* (cf. Theorems 3.2 and 3.3). The MMR allows us to develop a class of consistent CM tests that we call kernel conditional moment (KCM) tests (cf. Section 4). Furthermore, it considerably simplifies the parameter estimation problems based on the CMR. Our framework has relationships to existing methods in econometrics and machine learning (cf. Section 5). To the best of our knowledge, the Hilbert space embedding of conditional moment restrictions has not appeared elsewhere in the literature.[1]

All proofs can be found in Appendix D. The code of our experiments is available at `https://github.com/krikamol/kcm-test`.

2 BACKGROUND

We introduce the CMR in Section 2.1 and then review the concepts of kernels and RKHS in Section 2.2. Finally, we discuss the main assumptions in Section 2.3.

2.1 CONDITIONAL MOMENT RESTRICTIONS

Let Z be a random variable taking values in $\mathcal{Z} \subseteq \mathbb{R}^p$ with distribution P_Z, X a subvector of Z taking values in $\mathcal{X} \subseteq \mathbb{R}^d$ with distribution P_X, and $\Theta \subset \mathbb{R}^r$ a parameter space. Following Newey [38], we consider models where the only available information about the unknown parameter $\theta_0 \in \Theta$ is a set of conditional moment restrictions

$$\mathscr{M}(X;\theta_0) := \mathbb{E}[\boldsymbol{\psi}(Z;\theta_0)|X] = \mathbf{0}, \quad P_X\text{-a.s.}, \qquad (1)$$

where $\boldsymbol{\psi} : \mathcal{Z}\times\Theta \to \mathbb{R}^q$ is a vector of *generalized residual functions* whose functional forms are known up to the parameter $\theta \in \Theta$. The expectation is always taken over all random variables that are not conditioned on. Note that there can be two different models that are *observationally equivalent* on the basis of (1) alone although an ideal parameter θ_0 is unique.

Several statistical problems can be formulated as (1). In nonparametric regression models, $Z = (X, Y)$ where $Y \in \mathbb{R}$ is a dependent variable and $\boldsymbol{\psi}(Z;\theta) = Y - f(X;\theta)$. For conditional quantile models, $Z = (X, Y)$ and $\boldsymbol{\psi}(Z;\theta) = \mathbb{1}\{Y < f(X;\theta)\} - \tau$ for the target quantile $\tau \in [0, 1]$. In heterogeneous effect estimation, $Z = (X, T, Y)$ where T is a vector of treatments and $\boldsymbol{\psi}(Z;\theta(X)) = (Y - \langle\theta(X), T\rangle)T$. For instrumental variable regression, $Z = (X, W, Y)$ where W is an instrumental variable and $\boldsymbol{\psi}(Z;\theta) = (Y - f_\theta(X))$ and $\mathbb{E}[\boldsymbol{\psi}(Z;\theta)|W] = 0$ almost surely. When Z admits the density $p(z;\theta)$, we can define the moment conditions in terms of the *score function* as $\boldsymbol{\psi}(Z;\theta) = \nabla_\theta \log p(Z;\theta)$ and use it for local maximum likelihood estimation.

Conditional moment tests. Given an independent sample $(x_i, z_i)_{i=1}^n$ drawn from a distribution that satisfies the conditional moments (1) and an estimate $\hat{\theta}$ of θ_0, our goal is to perform specification testing: *Given a function $\boldsymbol{\psi}$ and a parameter estimate $\hat{\theta}$, we test the null hypothesis*

$$H_0 : \ \mathbb{E}[\boldsymbol{\psi}(Z;\hat{\theta}) \,|\, X] = \mathbf{0}, \ \ P_X\text{-a.s.} \qquad (2)$$

For instance, in the test of functional form of the nonlinear regression model [25], the null hypothesis can be expressed as $H_0 : \mathbb{E}[Y - f(X;\hat{\theta})|X] = 0$ where $\hat{\theta} = \arg\min_{\theta\in\Theta} \mathbb{E}[(Y - f(X;\theta))^2]$. In this case, $Z = (Y, X)$ and $\boldsymbol{\psi}(Z;\theta) = Y - f(X;\theta)$. This test allows us to detect misspecifications of the functional form of f.

In this work, we assume that $\hat{\theta}$ is obtained independently of the data that is used to test (2). In many cases, however, $\hat{\theta}$ is estimated using this data and hence the test performance is also subject to the estimation error. A generalization of our framework to those cases will require more involved analyses, and we leave it to future work.

[1] Carrasco and Florens [12] and their follow-up work are the most relevant works from the econometric literature. We discuss this connection in Section 5.

2.2 REPRODUCING KERNELS

Let $\mathcal{X}$ be a non-empty set and $\mathcal{F}$ a Hilbert space consisting of functions on $\mathcal{X}$ with $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ and $\|\cdot\|_{\mathcal{F}}$ being its inner product and norm, respectively. The Hilbert space $\mathcal{F}$ is called a reproducing kernel Hilbert space (RKHS) if there exists a symmetric function $k : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ called the reproducing kernel of $\mathcal{F}$ such that (i) $k(x,\cdot)\in\mathcal{F}$ for all $x\in\mathcal{X}$ and (ii) $f(x)=\langle f,k(x,\cdot)\rangle_{\mathcal{F}}$ for all $f\in\mathcal{F}$ and $x\in\mathcal{X}$. The latter is called the *reproducing property* of $\mathcal{F}$. Every positive definite kernel k uniquely determines the RKHS for which k is a reproducing kernel [5].

Let $\{(\lambda_j,e_j)\}$ be pairs of positive eigenvalues and orthonormal eigenfunctions of k, *i.e.*, $\int e_i(x)e_j(x)\,\mathrm{d}x = 1$ if $i=j$ and zero otherwise. By Mercer's theorem [49, Thm 4.49], the kernel k has the spectral decomposition

$$k(x,x') = \sum_j \lambda_j e_j(x)e_j(x'), \quad x,x'\in\mathcal{X}, \tag{3}$$

where the convergence is absolute and uniform. As a result, for any $f\in\mathcal{F}$, we have $f(x)=\sum_j f_je_j(x)$ with $\sum_j f_j^2/\lambda_j<\infty$ where $f_j=\langle f,e_j\rangle_{\mathcal{F}}$, $\langle f,g\rangle_{\mathcal{F}} = \sum_j f_jg_j/\lambda_j$, and $\|f\|^2_{\mathcal{F}}=\langle f,f\rangle_{\mathcal{F}}=\sum_j f_j^2/\lambda_j$.

Next, we introduce the notion of integrally strictly positive definite (ISPD) kernels and Bochner's characterization.

Definition 2.1. *A kernel $k(x,x')$ is integrally strictly positive definite (ISPD) if for any function f that satisfies $0<\|f\|_2^2<\infty$,*

$$\int_{\mathcal{X}} f(x)k(x,x')f(x')\,\mathrm{d}x\,\mathrm{d}x' > 0.$$

ISPD kernels are an important notion in kernel methods and are closely related to characteristic and universal kernels, see, *e.g.*, Simon-Gabriel and Schölkopf [45].

The next result characterizes shift-invariant kernels $k(x,x')=\varphi(x-x')$ for some positive definite φ.

Theorem 2.1 (Bochner)**.** *A continuous function $\varphi:\mathbb{R}^d\to\mathbb{C}$ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure Λ on $\mathbb{R}^d$:*

$$\varphi(t)=\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{-it^\top\omega}\,\mathrm{d}\Lambda(\omega)$$

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

Examples of popular kernels are the Gaussian RBF kernel $k(x,x')=\exp(-\|x-x'\|_2^2/2\sigma^2),\sigma>0$, Laplacian kernel $k(x,x')=\exp(-\|x-x'\|_1/\sigma),\sigma>0$, and inverse multiquadric (IMQ) kernel $k(x,x')=(c^2+\|x-x'\|_2^2)^{-\gamma}$, $c,\gamma>0$. See, *e.g.*, Steinwart and Christmann [49, Ch. 4] for more examples.

2.3 MAIN ASSUMPTIONS

Our subsequent analyses rely on these key assumptions.

- **(A1)** The random vector (X,Z) forms a strictly stationary process with the probability measure P_{XZ}.
- **(A2)** *Regularity conditions*: (i) the function $\boldsymbol{\psi}:\mathcal{Z}\times\Theta\to\mathbb{R}^q$ where $q<\infty$ is continuous on Θ for each $z\in\mathcal{Z}$; (ii) $\mathbb{E}[\boldsymbol{\psi}(Z;\theta)|x]$ exists and is finite for every $\theta\in\Theta$ and $x\in\mathcal{X}$ for which $P_X(x)>0$; (iii) $\mathbb{E}[\boldsymbol{\psi}(Z;\theta)|x]$ is continuous on Θ for all $x\in\mathcal{X}$ for which $P_X(x)>0$.
- **(A3)** *Global identification*: there exists a unique $\theta_0\in\Theta$ for which $\mathbb{E}[\boldsymbol{\psi}(Z;\theta_0)|X]=\mathbf{0}$ a.s., and $P(\mathbb{E}[\boldsymbol{\psi}(Z;\theta)|X]=\mathbf{0})<1$ for all $\theta\in\Theta,\theta\neq\theta_0$.
- **(A4)** The kernel k is ISPD, continuous, and bounded, *i.e.*, $\sup_{x\in\mathcal{X}}\sqrt{k(x,x)}<\infty$.

Assumption **(A1)** ensures that all expectations of functions of (X,Z) are independent of time. The regularity conditions **(A2)** are standard assumptions [22, Ch. 3] which ensure that $\boldsymbol{\psi}$ is well-defined, and hold in most models considered in the literature [22]. By contrast, **(A3)** may not hold, especially in non-linear models. A *local* identifiability can be assumed instead by imposing additional constraints on Θ. Testing whether the constraints are sufficient can then be done, for example, by examining the Jacobian at some parameter values [22, pp. 54]. Lastly, **(A4)** implies that the RKHS $\mathcal{F}$ consists of bounded continuous functions [49, Sec. 4.3] and is expressive enough (cf. Theorem 3.2).

3 MAXIMUM MOMENT RESTRICTION

This section presents the RKHS representation of the CMR. Let $\mathscr{F}$ be a set of measurable functions on $\mathcal{X}$. Then, $\mathbb{E}_{XZ}[\boldsymbol{\psi}(Z;\theta)f(X)] = \mathbb{E}_X[\mathbb{E}_Z[\boldsymbol{\psi}(Z;\theta)f(X)|X]] = \mathbb{E}_X[\mathscr{M}(X;\theta)f(X)]$ for any $f\in\mathscr{F}$ by the law of iterated expectation. That is, the CMR in (1) implies an infinite set of unconditional moment restrictions

$$\mathbb{E}[\boldsymbol{\psi}(Z;\theta_0)f(X)]=\mathbf{0}, \quad \forall f\in\mathscr{F}. \tag{4}$$

Equivalently, any $\theta_0\in\Theta$ that satisfies (4) must also satisfy what we call a *maximum moment restriction* (MMR)

$$\sup_{f\in\mathscr{F}} \|\mathbb{E}[\boldsymbol{\psi}(Z;\theta_0)f(X)]\|_2^2 = 0. \tag{5}$$

It is known that the implied moment restrictions (4) and (5) can be insufficient to globally identify the parameters of interest. We call $\mathscr{F}$ for which (5) implies (1) a *sufficient class of instruments*. In the context of this work, $\mathscr{F}$ must consist of infinitely many instruments for the CM test to be consistent against all alternatives. However, the $\sup$ operator also makes it hard to optimize (5). We resolve these issues by choosing $\mathscr{F}$ to be a unit ball in a RKHS, which we show to be a sufficient class of instruments. As a result, (5) can be solved analytically, the parameters of interest can be consistently estimated, and the resulting CM test is consistent against all fixed alternatives.

Recently, Lewis and Syrgkanis [30] and Bennett et al. [7] also propose to estimate θ_0 based on (5) and $\mathscr{F}$ that is parameterized by deep neural networks. While they consider an estimation problem, we focus on hypothesis testing problems. Nevertheless, our formulation of CMR can also be used to estimate θ_0 (cf. Section 3.2 and Appendix B). Note that the algorithms proposed in Lewis and Syrgkanis [30] and Bennett et al. [7] require solving a minimax game, whereas our approach for estimation is simply a minimization problem.

3.1 CONDITIONAL MOMENT EMBEDDING

To express (5) using the RKHS, we first develop a representation of the CMR in a vector-valued RKHS of functions $f : \mathcal{X} \to \mathbb{R}^q$ [2]. Let $\mathcal{F}$ be the RKHS of real-valued functions on $\mathcal{X}$ with reproducing kernel k and $\mathcal{F}^q$ the product RKHS of functions $f := (f_1, \dots, f_q)$ where $f_i \in \mathcal{F}$ for all i with an inner product $\langle f, g\rangle_{\mathcal{F}^q} = \sum_{i=1}^q \langle f_i, g_i\rangle_{\mathcal{F}}$ and norm $\|f\|_{\mathcal{F}^q} = \sqrt{\sum_{i=1}^q \|f_i\|^2_{\mathcal{F}}}$. For $\theta \in \Theta$, we define an operator M_θ on $\mathcal{F}^q$ as

$$M_\theta f := \mathbb{E}[\boldsymbol{\psi}(Z;\theta)^\top f(X)] = \sum_{i=1}^q \mathbb{E}[\psi_i(Z;\theta) f_i(X)],$$

where ψ_i denotes the i-th component of $\boldsymbol{\psi}$. This operator takes an instrument $f \in \mathcal{F}^q$ as input and returns the corresponding conditional moment restrictions.

The following lemma shows that M_θ satisfies the property of the original conditional moment restrictions.

Lemma 3.1. *For all $f \in \mathcal{F}^q$, $M_{\theta_0} f = 0$.*

Moreover, by Assumption **(A2)** and **(A4)**, $|M_\theta f| \leq \sum_{i=1}^q \|f_i\|_{\mathcal{F}_i} \sqrt{\mathbb{E}[\psi_i(Z;\theta)\psi_i(Z';\theta)k(X,X')]} < \infty$ where (X', Z') is an independent copy of (X, Z). Hence, M_θ is a bounded linear operator. By Riesz's representation theorem, there exists a unique element $\boldsymbol{\mu}_\theta$ in $\mathcal{F}^q$ such that $M_\theta f = \langle f, \boldsymbol{\mu}_\theta\rangle_{\mathcal{F}^q}$ for all $f \in \mathcal{F}^q$. Indeed, by the reproducing property,

$$M_\theta f = \sum_{i=1}^q \left\langle f_i, \mathbb{E}[\xi^i_\theta(X,Z)]\right\rangle_{\mathcal{F}_i} = \left\langle f, \mathbb{E}[\boldsymbol{\xi}_\theta(X,Z)]\right\rangle_{\mathcal{F}^q},$$

where $\boldsymbol{\xi}_\theta(x,z) := (\psi_1(z;\theta)k(x,\cdot), \dots, \psi_q(z;\theta)k(x,\cdot))$ is the feature map in $\mathcal{F}^q$ and ξ^i_θ denotes the i-th element of $\boldsymbol{\xi}_\theta$. The equalities above are well-defined since $\boldsymbol{\xi}_\theta(x,z)$ is Bochner integrable [49, Def. A.5.20], *i.e.*, $\mathbb{E}\|\boldsymbol{\xi}_\theta(X,Z)\|_{\mathcal{F}^p} \leq \sqrt{\mathbb{E}\|\boldsymbol{\xi}_\theta(X,Z)\|^2_{\mathcal{F}^p}} = \sqrt{\mathbb{E}[\boldsymbol{\psi}(Z;\theta)^\top \boldsymbol{\psi}(Z;\theta)k(X,X)]} < \infty$.

In other words, $\boldsymbol{\mu}_\theta := \mathbb{E}[\boldsymbol{\xi}_\theta(X,Z)]$ is a *representer* of M_θ in $\mathcal{F}^q$. We define $\boldsymbol{\mu}_\theta$ as *conditional moment embedding* (CMME) of $\mathbb{E}[\boldsymbol{\psi}(Z;\theta)|X]$ in $\mathcal{F}^q$ relative to P_X.

Definition 3.1. *For each $\theta \in \Theta$, let $\boldsymbol{\xi}_\theta(x,z) := (\psi_1(z;\theta)k(x,\cdot), \dots, \psi_q(z;\theta)k(x,\cdot)) \in \mathcal{F}^q$. The* conditional moment embedding *(CMME) is defined as*

$$\boldsymbol{\mu}_\theta := \int_{\mathcal{X}\times\mathcal{Z}} \boldsymbol{\xi}_\theta(x,z)\, dP_{XZ}(x,z) \in \mathcal{F}^q. \qquad (6)$$

4The CMME $\boldsymbol{\mu}_\theta$ takes the form of a kernel mean embedding of P_{XZ} with $\boldsymbol{\xi}_\theta$ as the feature map [33]. This is illustrated in Figure 1. Hence, given an i.i.d. sample $(x_i, z_i)_{i=1}^n$ from P_{XZ}, we can estimate $\boldsymbol{\mu}_\theta$ simply by $\widehat{\boldsymbol{\mu}}_\theta := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_\theta(x_i, z_i)$. The following theorem establishes the $\sqrt{n}$-consistency of this estimator.

Theorem 3.1. *Let $\sigma^2_\theta := \mathbb{E}\|\boldsymbol{\xi}_\theta(X,Z)\|^2_{\mathcal{F}^q}$ and assume that $\|\boldsymbol{\xi}_\theta(X,Z)\|_{\mathcal{F}^q} < C_\theta < \infty$ almost surely. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\|\widehat{\boldsymbol{\mu}}_\theta - \boldsymbol{\mu}_\theta\|_{\mathcal{F}^p} \leq \frac{2C_\theta \log\frac{2}{\delta}}{n} + \sqrt{\frac{2\sigma^2_\theta \log\frac{2}{\delta}}{n}}. \qquad (7)$$

Remarkably, $\widehat{\boldsymbol{\mu}}_\theta$ converges at a rate $O_p(n^{-1/2})$ that is independent of the dimension of (X, Z) and the RKHS $\mathcal{F}^q$. This is an appealing property because estimation and inference based on $\widehat{\boldsymbol{\mu}}_\theta$ become less susceptible to the *curse of dimensionality* (see, *e.g.*, Khosravi et al. [27] and references therein for the discussion). Under certain assumptions, Tolstikhin et al. [53] established the minimax optimal rate for the kernel mean estimators like $\widehat{\boldsymbol{\mu}}_\theta$.

The next theorem shows that $\boldsymbol{\mu}_\theta$ provides a *unique* representation of the CMR $\mathscr{M}(X, \theta)$ in $\mathcal{F}^q$ relative to P_X.

Theorem 3.2. *Assume that the kernel k is ISPD. Then, for any $\theta_1, \theta_2 \in \Theta$, $\mathscr{M}(x;\theta_1) = \mathscr{M}(x;\theta_2)$ for P_X-almost all x if and only if $\boldsymbol{\mu}_{\theta_1} = \boldsymbol{\mu}_{\theta_2}$.*

To better understand Theorem 3.2, consider when $q = 1$ and $k(x,x') = \varphi(x - x')$ is a shift-invariant kernel. First, we have $\boldsymbol{\mu}_\theta(\cdot) = \mathbb{E}_X[\mathbb{E}_Z[\boldsymbol{\psi}(Z;\theta)k(X,\cdot)|X]] = \mathbb{E}_X[\mathbb{E}_Z[\boldsymbol{\psi}(Z;\theta)|X]k(X,\cdot)] = \mathbb{E}_X[\mathscr{M}(X;\theta)k(X,\cdot)]$. It is then easy to show using Theorem 2.1 that $\boldsymbol{\mu}_\theta(\cdot) = \int_{\mathbb{R}^d} \phi(\omega;\theta)c(\omega,\cdot)\,\mathrm{d}\Lambda(\omega)$ where $c(\omega,y) = \exp(i\omega^\top y) \neq 0$ and $\phi(\omega;\theta) := \mathbb{E}_X[\mathscr{M}(X;\theta)\exp(i\omega^\top X)]$ is the Fourier transform (or characteristic function) of the Borel measurable function $\mathscr{M}(x;\theta)$ relative to P_X. Hence, if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, the uniqueness of $\boldsymbol{\mu}_\theta$ follows from the uniqueness of $\phi(\omega;\theta)$. Bierens [8] was the first to observe the characterization of the CMR in terms of the integral transform and then used it to construct the consistent CM tests of functional form (cf. Section 5).

Theorem 3.2 shows that $\boldsymbol{\mu}_\theta$ captures all information about $\mathbb{E}[\boldsymbol{\psi}(Z;\theta)|x]$ for every $x \in \mathcal{X}$ for which $P_X(x) > 0$. Consequently, estimation and inference on CMR can be performed by means of $\boldsymbol{\mu}_\theta$ using the existing kernel arsenal. As mentioned earlier, for each $f \in \mathcal{F}^q$ and $\theta \in \Theta$, the inner product $\langle f, \boldsymbol{\mu}_\theta \rangle_{\mathcal{F}^q} = \langle f, \mathbb{E}[\boldsymbol{\xi}_\theta(X,Z)]\rangle_{\mathcal{F}^q}$ can be interpreted as a restriction of conditional moments with respect to f. Moreover, the investigator can inspect $\boldsymbol{\mu}_\theta(x,z)$, which measures to what extent the moment conditions are violated at (x,z), *i.e.*, structural instability, in order to understand the nature of misspecification.

3.2 MAXIMUM MOMENT RESTRICTION WITH REPRODUCING KERNELS

Based on the CMME $\boldsymbol{\mu}_\theta$, we can now define the MMR as

$$\mathbb{M}(\theta) := \sup_{\|f\|_{\mathcal{F}^q} \leq 1} M_\theta f = \sup_{\|f\|_{\mathcal{F}^q}\leq 1} \langle f, \boldsymbol{\mu}_\theta \rangle_{\mathcal{F}^q} = \|\boldsymbol{\mu}_\theta\|_{\mathcal{F}^q}. \tag{8}$$

By Theorem 3.2, $\mathbb{M}(\theta) \geq 0$ and $\mathbb{M}(\theta) = 0$ if and only if $\theta = \theta_0$. Put differently, $\mathbb{M}(\theta)$ measures how much the models associated with θ violate the original CMR in (1).

To obtain an expression for $\mathbb{M}(\theta)$, we define a real-valued kernel $h_\theta : (\mathcal{X}\times\mathcal{Z})\times(\mathcal{X}\times\mathcal{Z}) \to \mathbb{R}$ based on the feature map $\boldsymbol{\xi}_\theta : \mathcal{X}\times\mathcal{Z} \to \mathcal{F}^q$ as follows:

$$\begin{aligned} h_\theta((x,z),(x',z')) &:= \langle \boldsymbol{\xi}_\theta(x,z), \boldsymbol{\xi}_\theta(x',z')\rangle_{\mathcal{F}^q} \\ &= \boldsymbol{\psi}(z;\theta)^\top \boldsymbol{\psi}(z';\theta) k(x,x'). \end{aligned} \tag{9}$$

Then, a closed-form expression for $\mathbb{M}(\theta)$ in terms of the kernel h_θ follows straightforwardly.

Theorem 3.3. *Assume that $\mathbb{E}[h_\theta((X,Z),(X,Z))] < \infty$. Then, $\mathbb{M}^2(\theta) = \mathbb{E}[h_\theta((X,Z),(X',Z'))]$ where (X',Z') is independent copy of (X,Z) with the same distribution.*

Finally, Mercer's representation (3) of k allows us to interpret h_θ and $\mathbb{M}(\theta)$ in terms of a continuum of unconditional moment restrictions.

Theorem 3.4. *Let $\{(\lambda_j, e_j)\}$ be eigenvalue/eigenfunction pairs associated with the kernel k and $\boldsymbol{\zeta}_\theta^j(x,z) := (\psi_1(z;\theta)e_j(x),\dots,\psi_q(z;\theta)e_j(x))$. Then, for each $\theta \in \Theta$, $h_\theta((x,z),(x',z')) = \sum_j \lambda_j \boldsymbol{\zeta}_\theta^j(x,z)^\top \boldsymbol{\zeta}_\theta^j(x',z')$ and $\mathbb{M}^2(\theta) = \sum_j \lambda_j \|\mathbb{E}[\boldsymbol{\zeta}_\theta^j(X,Z)]\|_2^2$.*

That is, we can interpret $\mathbb{E}[\boldsymbol{\zeta}_\theta^j(X,Z)]$ as the UMR with e_j acting as an instrument. Moreover, $\mathbb{M}^2(\theta)$ can be viewed as a weighted sum of moment restrictions based on the sequence of weights and instruments $(\lambda_j, e_j)_j$. As a result, the CM test based on $\mathbb{M}^2(\theta)$ as a test statistic examines an infinite number of moment restrictions. Note that $(\lambda_j, e_j)_j$ are defined implicitly by the choice of k.

4 KERNEL CONDITIONAL MOMENT TEST WITH BOOTSTRAPPING

By virtue of Theorem 3.2, we can reformulate the CM testing problem (2) in terms of the MMR as

$$H_0 : \mathbb{M}^2(\theta) = 0, \quad H_1 : \mathbb{M}^2(\theta) \neq 0.$$

Given an i.i.d. sample $\{(x_i,z_i)\}_{i=1}^n$ from the distribution P_{XZ}, we consider the test statistic

$$\widehat{\mathbb{M}}_n^2(\theta) = \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} h_\theta((x_i,z_i),(x_j,z_j)), \tag{10}$$

which is in the form of U-statistics [44, Section 5]. Although there exist several potential estimators for $\mathbb{M}^2(\theta)$, we focus on (10) as it is a minimum-variance unbiased estimator with appealing asymptotic properties. Moreover, (10) also provides a basis for the estimation of θ_0 simply by minimizing $\widehat{\mathbb{M}}_n^2(\theta)$ with respect to $\theta \in \Theta$. Preliminary results on estimation are given in Appendix B.

Next, we characterize the asymptotic distributions of $\widehat{\mathbb{M}}_n^2(\theta)$ under the null and alternative hypotheses.

Theorem 4.1. *Assume that $\mathbb{E}[h_\theta^2((X,Z),(X',Z'))] < \infty$ for all $\theta \in \Theta$. Let $U := (X,Z)$ and $U' := (X',Z')$. Then, the following statements hold.*

(1) *If $\theta \neq \theta_0$, $\widehat{\mathbb{M}}_n^2(\theta)$ is asymptotically normal with*

$$\sqrt{n}\left(\widehat{\mathbb{M}}_n^2(\theta) - \mathbb{M}^2(\theta)\right) \xrightarrow{d} \mathcal{N}(0, 4\sigma_\theta^2),$$

where $\sigma_\theta^2 = \mathrm{Var}_U\left[\mathbb{E}_{U'}[h_\theta(U,U')]\right]$.

(2) *If $\theta = \theta_0$, then $\sigma_\theta^2 = 0$ and*

$$n\widehat{\mathbb{M}}_n^2(\theta) \xrightarrow{d} \sum_{j=1}^\infty \tau_j \left(W_j^2 - 1\right), \tag{11}$$

where $W_j \sim \mathcal{N}(0,1)$ and $\{\tau_j\}$ are the eigenvalues of $h_\theta(u,u')$, i.e.*, they are the solutions of $\tau_j\phi_j(u) = \int h_\theta(u,u')\phi_j(u')dP(u')$ for non-zero ϕ_j.*

As we can see, $n\widehat{\mathbb{M}}_n^2(\theta) < \infty$ with probability one under the null $\theta = \theta_0$ and diverts to infinity at a rate $\mathcal{O}(\sqrt{n})$ under any fixed alternative $\theta \neq \theta_0$. Hence, a consistent CM test can be constructed as follows: if $\gamma_{1-\alpha}$ is the $1-\alpha$ quantile of the CDF of $n\widehat{\mathbb{M}}_n^2(\theta)$ under the null $\theta = \theta_0$, we reject the null with significance level α if $n\widehat{\mathbb{M}}_n^2(\theta) \geq \gamma_{1-\alpha}$.

Proposition 4.1 (Arcones and Gine [4]; p. 671)**.** *Assume the conditions of Theorem 4.1. The test that rejects the null $\theta = \theta_0$ when $n\widehat{\mathbb{M}}_n^2(\theta) > \gamma_{1-\alpha}$ is consistent against any fixed alternative $\theta \neq \theta_0$,* i.e.*, the limiting power of the test is one.*

Algorithm 1 KCM Test with bootstrapping

```
Input: Bootstrap sample size $B$, significance level $\alpha$
  for $t \in \{1, \dots, B\}$ do
    Draw $(w_1, \dots, w_n) \sim \mathrm{Mult}(n; \frac{1}{n}, \dots, \frac{1}{n})$
    $\rho_i \leftarrow (w_i - 1)/n$ for $i = 1, \dots, n$
    $\widehat{\mathbb{M}}^*_n(\theta) \leftarrow \sum_{i \neq j} \rho_i \rho_j h_\theta((x_i, z_i), (x_j, z_j))$
    $a_t \leftarrow n\widehat{\mathbb{M}}^*_n(\theta)$
  end for
  $\hat{\gamma}_{1-\alpha}$ := empirical $(1-\alpha)$-quantile of $\{a_t\}_{t=1}^B$
  Reject $H_0$ if $\hat{\gamma}_{1-\alpha} < n\widehat{\mathbb{M}}^2_n(\theta)$ (see (10))
```

Unfortunately, the limiting distribution in (11) and its $1-\alpha$ quantile do not have an analytic form. Following recent work on kernel-based tests [14, 21, 32], we propose to approximate the critical values using the bootstrap method proposed by Arcones and Gine [4], Huskova and Janssen [26], which was previously used in Liu et al. [32]. Specifically, we first draw multinomial random weights $(w_1, \dots, w_n) \sim \mathrm{Mult}(n; \frac{1}{n}, \dots, \frac{1}{n})$ and compute the bootstrap sample $\widehat{\mathbb{M}}^*_n(\theta) = (1/n^2)\sum_{1 \le i \neq j \le n}(w_i - 1)(w_j - 1)h_\theta((x_i, z_i), (x_j, z_j))$. We then calculate the empirical quantile $\hat{\gamma}_{1-\alpha}$ of $n\widehat{\mathbb{M}}^*_n(\theta)$. For degenerate U-statistics, $\hat{\gamma}_{1-\alpha}$ is a consistent estimate of $\gamma_{1-\alpha}$ [4, 26].

We summarize our bootstrap kernel conditional moment (KCM) test in Algorithm 1. Note that the proposed test checks the CMR for a *given* parameter θ and does not take into account the estimation error of θ. We defer a full treatment of interplay between parameter estimation and hypothesis testing to future work.

5 RELATED WORK

Existing CM tests can generally be categorized into two classes. The former is based on a transformation of CMR into a continuum of unconditional counterparts, *e.g.*, Bierens [8, 9], de Jong [15], Bierens and Ploberger [11], and Donald et al. [18] to name a few. The latter employs nonparametric kernel estimation which includes Fan and Li [20], Li and Wang [31], Zheng [56] among others. While both classes lead to consistent tests, they exhibit different asymptotic behaviors; see, *e.g.*, Delgado et al. [16], Fan and Li [20] for detailed comparisons.

A continuum of unconditional moments. One of the classical approaches is to find a parametric weighting function $w(x, \eta)$ such that

$$\mathbb{E}[\boldsymbol{\psi}(Z;\theta)|X] = \mathbf{0} \text{ a.s.} \ \Leftrightarrow\ \mathbb{E}[\boldsymbol{\psi}(Z;\theta)w(X,\eta)] = \mathbf{0},$$

for almost all $\eta \in \Xi \subseteq \mathbb{R}^m$ where η is a nuisance parameter. Newey [36] and Tauchen [52] proposed the so-called M-test using a finite number of weighting functions. Since it imposes only a finite number of moment conditions, the test cannot be consistent against all possible alternatives and power against specific alternatives depends on the choice of these weighting functions. de Jong [15] and Donald et al. [18] showed that this issue can be circumvented by allowing the number of moment conditions to grow with sample size. Although our KCM test generally relies on infinitely many moment conditions, one can impose finitely many conditions using the finite dimensional RKHS such as those endowed with linear and polynomial kernels or resorting to finite-dimensional kernel approximations.

Stinchcombe and White [50] showed that there exists a wide range of $w(x, \eta)$ that lead to consistent CM tests. They call these functions "totally revealing". For instance, Bierens [8] proposed the first consistent specification test for nonlinear regression models using $w(x, \eta) = \exp(i\eta^\top x)$ for $\eta \in \mathbb{R}^d$. Similarly, Bierens [9] used $w(x, \eta) = \exp(\eta^\top x)$ for $\eta \in \mathbb{R}^d$. An indicator function $w(x, \eta) = \mathbb{1}(\alpha^\top x \le \beta)$ with $\eta = (\alpha, \beta) \in \mathbb{S}^d \times (-\infty, \infty)$ where $\mathbb{S}^d = \{\alpha \in \mathbb{R}^d : \|\alpha\| = 1\}$ was used in Escanciano [19] and Delgado et al. [16]. Other popular weighting functions include power series, Fourier series, splines, and orthogonal polynomials, for example. In light of Theorem 3.4, the KCM test falls into this category where weighting functions are eigenfunctions associated with the kernel k.

Since $w(x, \eta)$ depends on the nuisance parameter η, Bierens [8] suggested to integrate η out, resulting in an *integrated conditional moment* (ICM) test statistic:

$$\widehat{T}_n(\theta) = \int_\Xi \|\widehat{Z}_n(\eta)\|_2^2 \,\mathrm{d}\nu(\eta), \tag{12}$$

where Ξ is a compact subset of $\mathbb{R}^d$, $\nu(\eta)$ is a probability measure on Ξ, and $\widehat{Z}_n(\eta) := (1/\sqrt{n})\sum_i \boldsymbol{\psi}(z_i;\theta)w(x_i,\eta)$. The limiting null distribution of the ICM test was proven to be a zero-mean Gaussian process [9]. Bierens and Ploberger [11] also characterizes the asymptotic null distribution of a general class of real-valued weighting functions.

The following theorem establishes the connection between the KCM and ICM test statistics.

Theorem 5.1. *Let $k(x, x') = \varphi(x - x')$ be a shift-invariant kernel on $\mathbb{R}^d$. Then, we have*

$$\mathbb{M}^2(\theta) = \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} \left\|\mathbb{E}[\boldsymbol{\psi}(Z;\theta)\exp(i\omega^\top X)]\right\|_2^2 \mathrm{d}\Lambda(\omega)$$

where Λ is a Fourier transform of k.

This theorem is quite insightful as it describes the KCM test statistic as the ICM test statistic $\widehat{T}_n(\theta)$ of

Bierens [8] where the distribution on the nuisance parameter ω is a Fourier transform of the kernel. For instance, the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|_2^2/2\sigma^2)$ corresponds to the Gaussian density $\Lambda(\omega) = \exp(-\sigma^2\|\omega\|_2^2/2)$; see Muandet et al. [33, Table 2.1] for more examples. Note that both weighting functions and integrating measures are implicitly determined by the kernel k. Unlike ICM tests, KCM tests can be evaluated without solving the high-dimensional numerical integration (12) explicitly. Moreover, KCM tests can be easily generalized to $\mathcal{X}$ that is not necessarily a subset of $\mathbb{R}^d$.

Carrasco and Florens [12] also considers a similar setting that involves a continuum of moment conditions in RKHS. Their approach, however, differs significantly from ours. First, they consider a specific case where the Hilbert space is a set of square integrable functions of a scalar $t \in [0, T]$ with the unconditional moment conditions $\mathbb{E}[\boldsymbol{\psi}_t(X, \theta_0)] = \mathbf{0}$ for all $t \in [0, T]$. Second, their key question is to identify the optimal choice of weighting matrix in GMM. Third, estimation is actually based on a truncation of infinite moment conditions. Lastly, they also proposed the CM test similar to the ICM tests, but it can handle only the case with $Z \in \mathbb{R}$, while our test is applicable to any domain with a valid kernel.

Nonparametric kernel estimation. The second class of tests, known as *smooth tests* [20, 31, 56], adopts the statistic of the form

$$T(\theta) = \mathbb{E}[\boldsymbol{\psi}(Z;\theta)^\top \mathbb{E}[\boldsymbol{\psi}(Z;\theta)|X]f(X)]. \qquad (13)$$

Based on the kernel estimator of $\mathbb{E}[\boldsymbol{\psi}(Z;\theta)|X]f(X)$, the empirical estimate of (13) can be expressed as

$$\widehat{T}_n(\theta) = \frac{1}{n(n-1)h^d} \sum_{1 \leq i \neq j \leq n} \boldsymbol{\psi}(z_i;\theta)^\top \boldsymbol{\psi}(z_j;\theta) K_{ij} \qquad (14)$$

where $K_{ij} = K((x_i - x_j)/h)$, $K(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a normalized kernel function and h is a smoothing parameter. Here, we emphasize that existing smooth tests rely on the kernel density estimator (KDE) in which the kernel used is not necessarily a reproducing kernel. For the smooth test to be consistent, h must vanish as $n \to \infty$, whereas our KCM test is consistent even when the kernel is fixed. Nevertheless, if $K(\cdot)$ is a reproducing kernel, the test statistic $\widehat{T}_n(\theta)$ with a fixed smoothing parameter h resembles the KCM test statistic (10). In fact, Fan and Li [20] has shown that the ICM test is a special case of the kernel-based test with a fixed smoothing parameter. However, the critical drawback of the nonparametric kernel-based tests is that they have non-trivial power only against local alternatives that approach the null at a slower rate than $1/\sqrt{n}$, due to the slower rate of convergence of kernel density estimators, *i.e.*, $O((nh^{d/2})^{-1/2})$ as $h \to 0$ [20]. Moreover, these tests are susceptible to the curse of dimensionality.

Last but not least, the kernel estimator is also a key ingredient in empirical likelihood-based CM tests [17, 28, 54].

Kernelized Stein discrepancy (KSD). Stein's methods [48] are among the most popular techniques in statistics and machine learning. One notable example is the Stein discrepancy which aims to characterize complex, high-dimensional distribution $p(x) = \tilde{p}(x)/N$ with intractable normalization constant $N = \int \tilde{p}(x)\,\mathrm{d}x$ using a *Stein operator* $\mathcal{A}_p$ such that

$$p = q \quad \Leftrightarrow \quad \mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)] = 0, \ \ \forall f, \qquad (15)$$

where $\mathcal{A}_p f(x) := \nabla_x \log p(x) f(x) + \nabla_x f(x)$. Here, we assume for simplicity that $x \in \mathbb{R}$. The Stein operator $\mathcal{A}_p$ depends on the density p through its *score function* $s_p(x) := \nabla_x \log p(x) = \frac{\nabla_x p(x)}{p(x)}$, which is independent of N. When $p \neq q$, the expectation in (15) gives rise to a discrepancy

$$\begin{aligned} \mathbb{S}_f(p, q) &:= \mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)] \\ &= \mathbb{E}_{x\sim q}[(s_p(x) - s_q(x))f(x)]. \end{aligned} \qquad (16)$$

See, also, Liu et al. [32, Lemma 2.3]. The Stein discrepancy has led to numerous applications such as variance reduction [40] and goodness-of-fit testing [14, 32], among others.

Like (4), we can observe that (15) is indeed a set of unconditional moment conditions. To make an explicit connection between Stein discrepancy and CMR, we need to assume access to the probability densities. Let $\mathcal{P}_\Theta$ be a space of probability densities $p(z;\theta)$ such that $\theta \mapsto p(z;\theta)$ is injective. We choose $\boldsymbol{\psi}(z;\theta) = \nabla_z \log p(z;\theta) =: s_\theta(z)$ as the associated score function.[2] This yields the following CMR:

$$\mathbb{E}[\nabla_z \log p(Z;\theta_0) \mid X] = \mathbf{0}, \quad P_X\text{-a.s.} \qquad (17)$$

For any $\theta \in \Theta$, it follows that $\mathbb{E}[\boldsymbol{\psi}(Z;\theta)^\top f(X)] = \mathbb{E}[s_\theta(Z)^\top f(X) - s_{\theta_0}(Z)^\top f(X)] = \mathbb{E}[(s_\theta(Z) - s_{\theta_0}(Z))^\top f(X)] =: \Delta_f(\theta, \theta_0)$. While $\Delta_f(\theta, \theta_0)$ resembles the Stein discrepancy in (16), we highlight the key differences. First, this characterization requires that the model is correctly specified, *i.e.*, $p(z;\theta_0)$ is observationally indistinguishable from the underlying data distribution. Second, like the Stein discrepancy, it can be interpreted as the $f(x)$-weighted expectation of the score difference $s_\theta - s_{\theta_0}$. In contrast, the weighting function $f(x)$ in our setting depends only on X, which is a subvector of Z. We provide further discussion about this

[2] This differs from the standard definition of score function as $\nabla_\theta \log p(z|\theta)$ in the interpretation of maximum likelihood as generalized method of moments [22].

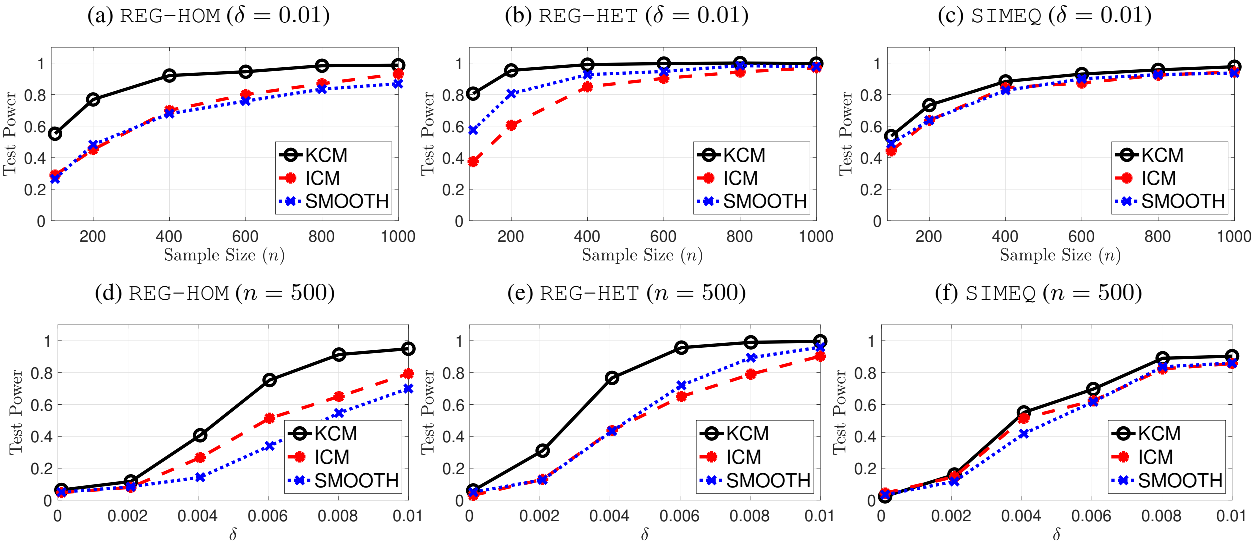


Figure 2: The test powers of KCM, ICM, and smooth tests averaged over 300 trials as we vary the values of n (top) and δ (bottom). Type-I errors of these tests are shown in Figure 3 in Appendix C.2. See main text for the interpretation.

discrepancy measure in Appendix A. The following theorem follows directly from the preceeding observation.

Theorem 5.2. *Let $\mathcal{P}_\Theta$ be a space of probability densities $p(z;\theta)$. Assume that $\theta \mapsto p(z;\theta)$ is injective and $\theta_0 \in \Theta$. If $\boldsymbol{\psi}(z;\theta) = \nabla_z \log p(z;\theta)$ and $X = Z$, we have $\mathbb{S}_f(p(z;\theta), p(z;\theta_0)) = \Delta_f(\theta,\theta_0)$.*

Mostly related to our work are the RKHS-based Stein's methods [14, 32]. Specifically, if we assume the conditions of Theorem 5.2 and that f belongs to the RKHS, it follows that $\Delta(\theta,\theta_0) := \sup_f \|\Delta_f(\theta,\theta_0)\|_2$ coincides with the kernelized Stein discrepancy (KSD) proposed in Liu et al. [32] and Chwialkowski et al. [14]. We will elaborate on this connection in further detail in future work.

6 EXPERIMENTS

We report the finite-sample performance of the KCM test against two well-known consistent CM tests, namely ICM test and smooth test, as discussed in Section 5. We evaluate all tests with a bootstrap size $B = 1000$ and a significance level $\alpha = 0.05$.

(1) **KCM**: The bootstrap KCM test using U-statistic in Algorithm 1. We use the RBF kernel with bandwidth chosen by the median heuristic.

(2) **ICM**: The test based on an integration over weighting functions. Following Stute [51] and Delgado et al. [16], we use (12) as the test statistic with $w(x,\eta) = \mathbb{1}(x \le \eta) = \prod_{j=1}^d \mathbb{1}(x_j \le \eta_j)$ where $\mathbb{1}(\cdot)$ is an indicator function. The density ν is chosen to be the empirical distribution of X. This leads to a simple test statistic $t_n = \sum_{i=1}^n r_n(x_i)^\top r_n(x_i)$ where $r_n(x) := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(z_i;\theta)\mathbb{1}(x_i \le x)$. We follow the bootstrap procedure in Delgado et al. [16, Sec. 4.3] to compute the critical values.

(3) **Smooth**: The test based on nonparametric kernel estimation. We use (14) as the test statistic. The kernel is the standard Gaussian density function whose bandwidth is chosen by the rule-of-thumb $h = n^{-1/5}$. Note that the median heuristic is not applicable here because the bandwidth h does not vanish, as required. The critical values are obtained using the same bootstrap procedure as in Delgado et al. [16, Sec. 4.2].

Testing a regression function (REG). We follow a similar simulation of regression model used in Lavergne and Nguimkeu [29]. In this setting, for a given estimate $\hat{\boldsymbol{\beta}}$ of the regression parameters, the null hypothesis is

$$H_0 : \mathbb{E}[Y - \hat{\boldsymbol{\beta}}^\top X \mid X] = 0 \quad \text{a.s.}$$

where $X \in \mathbb{R}^d$ and Y is a univariate random variable, *i.e.*, $Z = (Y, X)$. The data are generated from the data generating process (DGP):

$$Y = \boldsymbol{\beta}_0^\top X + e.$$

We set $\boldsymbol{\beta}_0 = \mathbf{1}$, and $X \sim \mathcal{N}(0, I_d)$. For the error term e, we consider two scenarios: (i) *Homoskedastic* (HOM): $e = \epsilon, \epsilon \sim \mathcal{N}(0,1)$ and (ii) *Heteroskedastic* (HET): $e = \epsilon\sqrt{0.1 + 0.1\|X\|_2^2}$. In each trial, we obtain an estimate of $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \gamma$ where $\gamma \sim \mathcal{N}(\mathbf{0}, \delta^2 I_d)$. In this experiment, we set $d = 5$. When $\delta = 0$, the CMR are fulfilled, whereas they are violated, *i.e.*, H_0 is false, if $\delta \neq 0$. Different values of δ correspond to different degrees of deviation from the null.

Testing the simultaneous equation model (SIMEQ). Following Newey [37] and Delgado et al. [16], we consider the equilibrium model

$$\begin{aligned} Q &= \alpha_d P + \beta_d R + U, \quad \alpha_d < 0, \qquad \text{(Demand)} \\ Q &= \alpha_s P + \beta_s W + V, \quad \alpha_s > 0, \qquad \text{(Supply)} \end{aligned}$$

where Q and P denote quantity and price, respectively, R and W are exogeneous variables, and U and V are the error terms. In this setting, $Z = (Q, P, R, W)$ and $X = (R, W)$. The null hypothesis can be expressed as

$$H_0 : \mathbb{E}\left[\begin{array}{c} Q - \alpha_d P - \beta_d R \\ Q - \alpha_s P - \beta_s W \end{array} \middle| \, \mathrm{X} \right] = \left[\begin{array}{c} 0 \\ 0 \end{array} \right]$$

a.s. for some $\theta_0 = (\alpha_d, \beta_d, \alpha_s, \beta_s)$. We generate data according to $Q = \lambda_{11}R + \lambda_{12}W + V_1$ and $P = \lambda_{21}R + \lambda_{22}W + V_2$ where R and W are independent standard Gaussian random variables while V_1 and V_2 are correlated standard Gaussian random variables with 10^{-3} variance and $10^{-3}/\sqrt{2}$ covariance, and independent of (R, W). We set $(\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}) = (1, -1, 1, 1)$ and provide the details on how to find the true parameters θ_0 in Appendix C.1. The estimate $\hat{\theta}$ is obtained as in the previous experiment. The null hypothesis corresponds to $\delta = 0$ and different values of δ corresponds to alternative hypotheses. Rejecting H_0 means that the functional form of the supply and demand curves are misspecified.

Figure 2 depicts the empirical results for $n \in \{1, 2, 4, 6, 8, 10\} \times 10^2$ and $\delta \in \{10^{-4}, 2 \times 10^{-3}, 4 \times 10^{-3}, 6 \times 10^{-3}, 8 \times 10^{-3}, 10^{-2}\}$. First, it can be observed that KCM, ICM, and smooth tests are all capable of detecting the misspecification as the sample size and δ are sufficiently large. Second, the KCM test tends to outperform both ICM and smooth tests in terms of the test power, especially in a low sample regime (see Figure 2a–2c) and a small deviation regime (see Figure 2d–2f). In addition, the smooth test and the ICM test are competitive: there is no substantial evidence to conclude that one is always better than the other. Lastly, Figure 3 in Appendix C.2 depicts that the Type-I errors of all tests are correctly controlled at $\alpha = 0.05$.

Lastly, we point out that this work does not elaborate on the effect of parameter estimation. In practice, the candidate parameter $\hat{\theta}$ has to be estimated from the observed data, which changes the asymptotic distribution of the test statistic. We envision the interplay between parameter estimation and hypothesis testing as an important arena for future work.

7 CONCLUSION

To conclude, we propose a new conditional moment test called the KCM test whose statistic is based on a novel representation of the conditional moment restrictions in a reproducing kernel Hilbert space. This representation captures all necessary information about the original conditional moment restrictions. Hence, the resulting test is consistent against all fixed alternatives, is easy to use in practice, and also has connections to existing tests in the literature. It also has an encouraging finite-sample performance compared to those tests. While the conditional moment restrictions have a long history in econometrics and so does the concept of reproducing kernel Hilbert spaces in machine learning, the intersection of these concepts remains unexplored. We believe that this work gives rise to a new and promising framework for conditional moment restrictions which constitute numerous applications in econometrics, causal inference, and machine learning.

Acknowledgements

We thank the anonymous reviewers for the helpful comments on our initial draft. KM is indebted to Motonobu Kanagawa and Junhyung Park for fruitful discussion.

References

[1] C. Ai and X. Chen. Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica*, 71(6):1795–1843, 2003.

[2] M. Álvarez, L. Rosasco, and N. Lawrence. Kernels for vector-valued functions: A review. *Foundation and Trends in Machine Learning*, 4(3):195–266, 2012.

[3] J. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

[4] M. Arcones and E. Gine. On the bootstrap of U and V statistics. *The Annals of Statistics*, 20(2):655–674, 06 1992.

[5] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[6] S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 04 2019.

[7] A. Bennett, N. Kallus, and T. Schnabel. Deep generalized method of moments for instrumental variable analysis. In *NeurIPS*, 2019.

[8] H. Bierens. Consistent model specification tests. *Journal of Econometrics*, 20(1):105–134, 1982.

[9] H. Bierens. A consistent conditional moment test of functional form. *Econometrica*, 58(6):1443–1458, 1990.

[10] H. Bierens. *Econometric Model Specification: Consistent Model Specification Tests and Semi-nonparametric Modeling and Inference*. World Scientific, 2017.

[11] H. Bierens and W. Ploberger. Asymptotic Theory of Integrated Conditional Moment Tests. *Econometrica*, 65(5):1129–1152, 1997.

[12] M. Carrasco and J.-P. Florens. Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 16(6):797–834, 2000.

[13] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[14] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *ICML*, 2016.

[15] R. de Jong. The Bierens test under data dependence. *Journal of Econometrics*, 72(1-2):1–32, 1996.

[16] M. Delgado, M. Domínguez, and P. Lavergne. Consistent tests of conditional moment restrictions. *Annales d'Économie et de Statistique*, (81):33–67, 2006.

[17] M. Dominguez and I. Lobato. Consistent estimation of models defined by conditional moment restrictions. *Econometrica*, 72(5):1601–1615, 2004.

[18] S. Donald, G. Imbens, and W. Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, 117(1):55–93, 2003.

[19] J. C. Escanciano. A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22 (6):1030–1051, 2006.

[20] Y. Fan and Q. Li. Consistent model specification tests: Kernel-based tests versus Bierens' ICM tests. *Econometric Theory*, 16:1016–1041, 12 2000.

[21] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[22] A. Hall. *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press, 2005.

[23] L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4): 1029–1054, 1982.

[24] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep IV: A flexible approach for counterfactual prediction. In *ICML*, 2017.

[25] J. Hausman. Specification tests in econometrics. *Econometrica*, 46(6):1251–71, 1978.

[26] M. Huskova and P. Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, 21(4):1811–1823, 12 1993.

[27] K. Khosravi, G. Lewis, and V. Syrgkanis. Nonparametric inference adaptive to intrinsic dimension. *ArXiv:1901.03719*, 2019.

[28] Y. Kitamura, G. Tripathi, and H. Ahn. Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72(6):1667–1714, 2004.

[29] P. Lavergne and P. Nguimkeu. A Hausman Specification Test of Conditional Moment Restrictions. TSE Working Papers 16-743, Toulouse School of Economics (TSE), 2016.

[30] G. Lewis and V. Syrgkanis. Adversarial generalized method of moments. *ArXiv:1803.07164*, 2018.

[31] Q. Li and S. Wang. A simple consistent bootstrap test for a parametric regression function. *Journal of Econometrics*, 87(1):145–165, 1998.

[32] Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML*, 2016.

[33] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

[34] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj. Dual instrumental variable regression. *ArXiv:1910.12358*, 2019.

[35] J. Muth. Rational expectations and the theory of price movements. *Econometrica*, 29(3):315–335, 1961.

[36] W. Newey. Maximum likelihood specification testing and conditional moment tests. *Econometrica*, 53(5):1047–1070, 1985.

[37] W. Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(4):809–837, 1990.

[38] W. Newey. Efficient estimation of models with conditional moment restrictions. In *Handbook of Statistics*, volume 11, chapter 16, pages 419–454. 1993.

[39] W. Newey and D. McFadden. Large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111–2245. Elsevier, 1994.

[40] C. Oates, M. Girolami, and N. Chopin. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society Series B*, 79(3):695–718, 2017.

[41] M. Oprescu, V. Syrgkanis, and Z. S. Wu. Orthogonal random forest for causal inference. In *ICML*, 2019.

[42] I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4): 1679–1706, 10 1994.

[43] J. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, 1958.

[44] R. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

[45] C.-J. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018.

[46] R. Singh, M. Sahani, and A. Gretton. Kernel instrumental variable regression. In *NeurIPS*, 2019.

[47] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 08 2007.

[48] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 583–602. University of California Press, 1972.

[49] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

[50] M. Stinchcombe and H. White. Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory*, 14(3):295–325, 1998.

[51] W. Stute. Nonparametric model checks for regression. *The Annals of Statistics*, 25(2):613–641, 04 1997.

[52] G. Tauchen. Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, 30(1): 415–443, 1985.

[53] I. Tolstikhin, B. Sriperumbudur, and K. Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18:86:1–86:47, 2017.

[54] G. Tripathi and Y. Kitamura. Testing conditional moment restrictions. *The Annals of Statistics*, 31(6):2059–2095, 2003.

[55] H. White. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, 76(374):419–433, 1981.

[56] J. Zheng. A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, 75(2):263–289, 1996.

Appendix

Table of Contents

A Conditional Moment Discrepancy (CMMD)

The maximum moment restriction (MMR) also allows us to compare two different models based on the conditional moment restriction (CMR). Let $\mathcal{M}_{\theta_1}$ and $\mathcal{M}_{\theta_2}$ be two models parameterized by $\theta_1, \theta_2 \in \Theta$, respectively. Then, we can define a CMR-based discrepancy measure between these two models as follows.

Definition A.1. *For $\theta_1, \theta_2 \in \Theta$, a conditional moment discrepancy (CMMD) is defined as $\Delta(\theta_1, \theta_2) := \|\boldsymbol{\mu}_{\theta_1} - \boldsymbol{\mu}_{\theta_2}\|_{\mathcal{F}^p}$.*

By Theorem 3.2, $\Delta(\theta_1, \theta_2) \geq 0$ and $\Delta(\theta_1, \theta_2) = 0$ if and only if the two models $\mathcal{M}_{\theta_1}$ and $\mathcal{M}_{\theta_2}$ are indistinguishable in terms of the CMR alone. Moreover, if the global identifiability **(A3)** holds, $\Delta(\theta_0, \theta) = \mathbb{M}(\theta)$ for all $\theta \in \Theta$. Since

$$\Delta(\theta_1, \theta_2) = \|\mathbb{E}[\boldsymbol{\xi}_{\theta_1}(X, Z) - \boldsymbol{\xi}_{\theta_2}(X, Z)]\|_{\mathcal{F}^q} = \left\|\mathbb{E}[\bar{\boldsymbol{\xi}}(X, Z)]\right\|_{\mathcal{F}^q}$$

where $\bar{\boldsymbol{\xi}}(x, z) := \boldsymbol{\xi}_{\theta_1}(x, z) - \boldsymbol{\xi}_{\theta_2}(x, z) = (\boldsymbol{\psi}(z; \theta_1) - \boldsymbol{\psi}(z; \theta_2))k(x, \cdot)$, the CMMD can be viewed as the MMR defined on a *differential residual function* $\boldsymbol{\psi}(z; \theta_1) - \boldsymbol{\psi}(z; \theta_2)$. As a result, $\Delta(\theta_1, \theta_2)$ also has a closed-form expression similar to that in Theorem 3.3.

Corollary A.1. *For $\theta_1, \theta_2 \in \Theta$, let*

$$h((x, z), (x', z')) := (\boldsymbol{\psi}(z; \theta_1) - \boldsymbol{\psi}(z; \theta_2))^\top (\boldsymbol{\psi}(z'; \theta_1) - \boldsymbol{\psi}(z'; \theta_2))k(x, x')$$

and assume that $\mathbb{E}[h((X, Z), (X, Z))] < \infty$. Then, we have $\Delta^2(\theta_1, \theta_2) = \mathbb{E}[h((X, Z), (X', Z'))]$ where (X', Z') is independent copy of (X, Z) with an identical distribution.

Proof. The result follows by applying the proof of Theorem 3.3 to the feature map $\bar{\boldsymbol{\xi}}(x, z) := \boldsymbol{\xi}_{\theta_1}(x, z) - \boldsymbol{\xi}_{\theta_2}(x, z) = (\boldsymbol{\psi}(z; \theta_1) - \boldsymbol{\psi}(z; \theta_2))k(x, \cdot)$. □

Furthermore, we can express the empirical CMMD as

$$\Delta_n^2(\theta_1,\theta_2) := \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} h((x_i,z_i),(x_j,z_j))$$

where $h((x_i,z_i),(x_j,z_j)) := (\boldsymbol{\psi}(z_i;\theta_1) - \boldsymbol{\psi}(z_i;\theta_2))^\top(\boldsymbol{\psi}(z_j;\theta_1) - \boldsymbol{\psi}(z_j;\theta_2))k(x_i,x_j)$.

As we can see, the RKHS norm, inner product, and function evaluation computed with respect to $\boldsymbol{\mu}_\theta$ all have meaningful economic interpretations. Table 1 summarizes these interpretations.

Table 1: Interpretations of different operations on $\boldsymbol{\mu}_\theta$ in $\mathcal{F}^q$.

Operation	**Interpretation**
$\|\boldsymbol{\mu}_\theta\|_{\mathcal{F}^q}$	conditional moment violation
$\langle f, \boldsymbol{\mu}_\theta\rangle_{\mathcal{F}^q}$	violation w.r.t. the instrument f
$\boldsymbol{\mu}_\theta(x,z)$	structural instability at (x,z)
$\|\boldsymbol{\mu}_{\theta_1} - \boldsymbol{\mu}_{\theta_2}\|_{\mathcal{F}^q}$	discrepancy between $\mathcal{M}_{\theta_1}$ and $\mathcal{M}_{\theta_2}$

B Parameter Estimation

Besides hypothesis testing, another important application of the CMR is parameter estimation. That is, given the CMR as in (1), we aim to find an estimate of θ_0 that satisfies (1) from the observed data $(x_i,z_i)_{i=1}^n$. Based on the MMR, we define the estimator of θ_0 as the parameter that minimizes (10):

$$\hat{\theta}_n := \arg\min_{\theta\in\Theta} \widehat{\mathbb{M}}_n^2(\theta) = \arg\min_{\theta\in\Theta} \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} h_\theta((x_i,z_i),(x_j,z_j)). \tag{18}$$

We call $\hat{\theta}_n$ a *minimum maximum moment restriction* (MMMR) estimate of θ_0. Note that it is also possible to adopt V-statistic in (18) instead of the U-statistic. Previously, Lewis and Syrgkanis [30] and Bennett et al. [7] proposed to estimate θ_0 based on (5) and $\mathscr{F}$ that is parameterized by deep neural networks. However, their algorithms require solving a minimax game, whereas our approach for estimation is merely a minimization problem.

The following theorem shows that $\hat{\theta}_n$ is a consistent estimate of θ_0. The proof can be found in Appendix D.6.

Theorem B.1 (Consistency of $\hat{\theta}_n$)**.** *Assume that the parameter space Θ is compact. Then, we have $\hat{\theta}_n \xrightarrow{p} \theta_0$.*

Despite the consistency, we suspect that $\hat{\theta}_n$ may not be asymptotically efficient and there exist better estimators. Theorem 3.4 shows that $\mathbb{M}(\theta)$ depends on a continuum of moment conditions reweighted by the non-uniform eigenvalues $(\lambda_j)_j$, which suggests that a *reweighting matrix* must also be incorporated in order to achieve the optimality [22]. Constructing an optimal choice of reweighting matrix in an infinite dimensional RKHS is an interesting topic [12], and we leave it to future work.

B.1 Maximum Moment Restriction for Instrumental Variable Regression

To illustrate one of the advantages of the MMR for parameter estimation, let us consider the nonparametric instrumental variable regression problem [3, 7, 30, 34, 46]. Let X be a treatment (endogeneous) variable taking values in $\mathcal{X} \subseteq \mathbb{R}^d$ and Y a real-valued outcome variable. Our goal is to estimate a function $g : \mathcal{X} \to \mathbb{R}$ from a structural equation model (SEM) of the form

$$Y = g(X) + \varepsilon, \quad X = h(Z) + f(\varepsilon) + \nu, \tag{19}$$

where we assume that $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\nu] = 0$. Unfortunately, as we can see from (19), ε is correlated with the treatment X, i.e., $\mathbb{E}[\varepsilon|X] \neq 0$, and hence standard regression methods cannot be used to estimate g. This setting often arises when there exist unobserved confounders between the treatment X and outcome Y.

In instrumental variable regression, we assume access to an *instrumental* variable Z which is associated with the treatments X, but not with the outcome variable Y, other than through its effect on the treatments. Moreover, the instrument Z is assumed to be uncorrelated with ε. This implies the conditional moment restriction $\mathbb{E}[\varepsilon \,|\, Z] = \mathbb{E}[Y - g(X) \,|\, Z] = 0$ for P_Z-almost all z [7, 30, 38]. Given an i.i.d. sample $(x_i, y_i, z_i)_{i=1}^n$ from $P(X, Y, Z)$, the MMR allows us to reduce the problem of estimating g to a regularized empirical risk minimization (ERM) problem

$$\begin{aligned} \widehat{g}_\lambda \;&:=\; \arg\min_{g\in\mathcal{G}_l} \widehat{\mathbb{M}}_n^2(g) + \lambda\|g\|_{\mathcal{G}_l}^2 \\ &=\; \arg\min_{g\in\mathcal{G}_l} \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (y_i - g(x_i))(y_j - g(x_j))k(z_i, z_j) + \lambda\|g\|_{\mathcal{G}_l}^2, \end{aligned} \tag{20}$$

where λ is a positive regularization parameter and $\mathcal{G}_l$ is a reproducing kernel Hilbert space (RKHS) of real-valued functions on $\mathcal{X}$ with the reproducing kernel $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Note that we adopt the V-statistic instead of the U-statistic in (20). By the representer theorem, the optimal solution to (20) can be expressed as a linear combination

$$\widehat{g}_\lambda(x) = \sum_{i=1}^n \alpha_i l(x, x_i) \tag{21}$$

for some $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. Let $K = [k(z_i, z_j)]_{i,j}$ and $L = [l(x_i, x_j)]_{i,j}$ be the kernel matrices in $\mathbb{R}^{n\times n}$ of $\boldsymbol{z} = [z_1, \ldots, z_n]^\top$ and $\boldsymbol{x} = [x_1, \ldots, x_n]^\top$, respectively, and $\boldsymbol{y} := [y_1, \ldots, y_n]^\top$. Substituting (21) back into (20) yields a *generalized ridge regression* (GRR) problem

$$\boldsymbol{\alpha}_\lambda \;:=\; \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \frac{1}{n^2}(\boldsymbol{y} - L\boldsymbol{\alpha})^\top K(\boldsymbol{y} - L\boldsymbol{\alpha}) + \lambda\boldsymbol{\alpha}^\top L\boldsymbol{\alpha}. \tag{22}$$

That is, the optimal coefficients $\boldsymbol{\alpha}_\lambda$ can be obtained by solving the first-order stationary condition $(LKL + n^2\lambda L)\boldsymbol{\alpha} = LK\boldsymbol{y}$ and if L is positive definite, the solution has a *closed-form* expression, *i.e.*,

$$\widehat{g}_\lambda(x) = \sum_{i=1}^n \alpha_{\lambda,i} l(x, x_i), \qquad \boldsymbol{\alpha}_\lambda = (LKL + n^2\lambda L)^{-1}LK\boldsymbol{y}. \tag{23}$$

Similar technique has been considered in Singh et al. [46] and Muandet et al. [34]. In Singh et al. [46], the authors extended the two-stage least square (2SLS) by modeling the first-stage regression with the conditional mean embedding of $P(X|Z)$ [33] which is then used in the second-stage kernel ridge regression. In Muandet et al. [34], the authors showed that the two-stage procedure can be reformulated as a convex-concave saddle-point problem. When the solutions lie in the RKHS, the closed-form solution similar to (23) and the one in Singh et al. [46] can be obtained. By contrast, the MMR-based approach allows us to reformulate the problem directly as a generalized ridge regression (GRR) in which the values of hyperparameters, *e.g.*, the regularization parameter λ, can be chosen via the popular cross-validation procedures.

C Experiments

In this section, we provide further description of our experiments as well as additional experimental results.

C.1 Simultaneous Equation Models

A simultaneous equation model (SEM) is a fundamental concept in economics. In one of our experiments, we consider the following SEM:

$$\begin{aligned} Q &= \alpha_d P + \beta_d R + U, \quad \alpha_d < 0, \qquad \text{(Demand)} \\ Q &= \alpha_s P + \beta_s W + V, \quad \alpha_s > 0, \qquad \text{(Supply)} \end{aligned} \tag{24}$$

where Q and P denote quantity and price, respectively, R and W are exogeneous variables, and U and V are the error terms. To obtain *reduced-form equations* of (24), we must solve for the endogeneous variables P and Q. First, we solve for P by equating the two equations in (24):

$$P = \left[\frac{\beta_s}{\alpha_d - \alpha_s}\right] W - \left[\frac{\beta_d}{\alpha_d - \alpha_s}\right] R + \frac{V - U}{\alpha_d - \alpha_s}. \tag{25}$$

Then, we can solve for Q by plugging in P to the supply equation in (24):

$$Q = \left[\frac{\alpha_s\beta_s}{\alpha_d - \alpha_s} + \beta_s\right] W - \left[\frac{\alpha_s\beta_d}{\alpha_d - \alpha_s}\right] R + \frac{\alpha_s}{\alpha_d - \alpha_s}(V - U) + V. \tag{26}$$

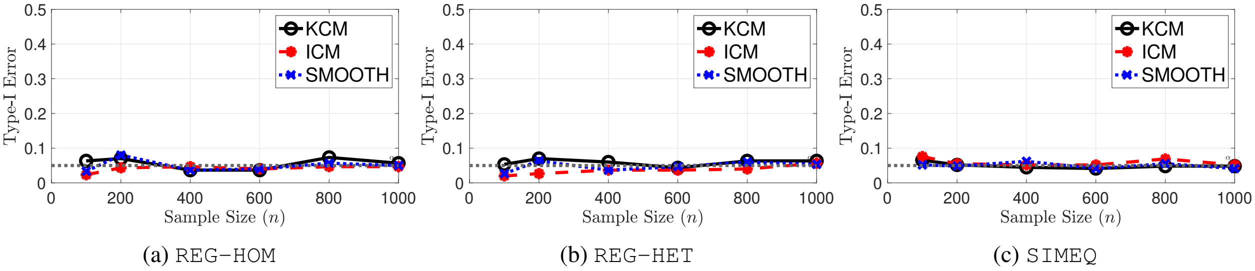
By comparing (25) and (26) to the data generating process in our experiment, we obtain the following system of equations:

$$\begin{aligned} \lambda_{11} &= -\frac{\alpha_s\beta_d}{\alpha_d - \alpha_s}, & \lambda_{21} &= -\frac{\beta_d}{\alpha_d - \alpha_s} \\ \lambda_{12} &= \frac{\alpha_s\beta_s}{\alpha_d - \alpha_s} + \beta_s, & \lambda_{22} &= \frac{\beta_s}{\alpha_d - \alpha_s}. \end{aligned} \tag{27}$$

Finally, setting $(\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}) = (1, -1, 1, 1)$ and then solving (27) result in a non-trivial solution $(\alpha_d, \beta_d, \alpha_s, \beta_s) = (-1, 2, 1, -2)$. This solution coincides with the one obtained from the two-stage least square (2SLS) procedure [3, Ch. 4].

C.2 Type-I Errors

The KCM test with bootstrapping is based on the asymptotic distribution of the test statistic under H_0 (cf. Theorem 4.1). Hence, the test reliably controls the Type-I error when the sample size is sufficiently large, *i.e.*, we are in the asymptotic regime. For the considered examples, this is the case already for moderate sample sizes. We report the Type-I error at a significance level $\alpha = 0.05$ for $n \in \{100, 200, 400, 600, 800, 1000\}$ in Figure 3.



(a) REG-HOM (b) REG-HET (c) SIMEQ

Figure 3: The Type-I errors averaged over 300 trials of KCM, ICM, and smooth tests under the null hypothesis ($\delta = 0$) as we vary the sample size n.

D Proofs

This section collects all the proofs of the results presented in the main paper.

D.1 Proof of Lemma 3.1

Proof. We have $M_{\theta_0} f = \sum_{i=1}^{q} \mathbb{E}[\psi_i(Z;\theta_0) f_i(X)]$ and, for all $i = 1, \dots, q$,

$$\mathbb{E}_{XZ}[\psi_i(Z;\theta_0) f_i(X)] = \mathbb{E}_X[\mathbb{E}_Z[\psi_i(Z;\theta_0) f_i(X)|X]] = \mathbb{E}_X[\mathbb{E}_Z[\psi_i(Z;\theta_0)|X] f_i(X)] = 0$$

by the law of iterated expectation. The last equality follows from the definition of θ_0 and the continuity of f_i, *i.e.*, by Assumption **(A4)**. □

D.2 Proof of Theorem 3.1

Our result follows directly from Smale and Zhou [47, Lemma 2] and Pinelis [42, Theorem 3.4] which rely on the Bennett inequality for vector-valued random variables. We reproduce the proof here for completeness.

Proof. First, recall that $\boldsymbol{\mu}_\theta = \mathbb{E}[\boldsymbol{\xi}_\theta(X,Z)]$ and $\widehat{\boldsymbol{\mu}}_\theta = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_\theta(x_i,z_i)$ for the independent random variables $\{\boldsymbol{\xi}_\theta(x_i,z_i)\}_{i=1}^n$. Then, for any $\varepsilon > 0$, it follows from Smale and Zhou [47, Lemma 1] that

$$P\left\{\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_\theta(x_i,z_i) - \boldsymbol{\mu}_\theta\right\|_{\mathcal{F}^q} \geq \varepsilon\right\} \leq 2\exp\left\{-\frac{n\varepsilon}{2C_\theta}\log\left(1+\frac{C_\theta\varepsilon}{\sigma_\theta^2}\right)\right\}.$$

Taking $t := C_\theta\varepsilon/\sigma_\theta^2$ and applying the inequality $\log(1+t) \geq t/(1+t)$ for all $t > 0$ yield

$$\begin{aligned} P\left\{\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_\theta(x_i,z_i) - \boldsymbol{\mu}_\theta\right\|_{\mathcal{F}^q} \geq \varepsilon\right\} &\leq 2\exp\left\{-\frac{n\varepsilon}{2C_\theta}\left(\frac{C_\theta\varepsilon}{C_\theta\varepsilon+\sigma_\theta^2}\right)\right\} \\ &= 2\exp\left\{-\frac{n\varepsilon^2}{2C_\theta\varepsilon+2\sigma_\theta^2}\right\}. \end{aligned}$$

The value of $\varepsilon > 0$ for which this probability equal to δ can be obtained by solving the quadratic equation $n\varepsilon^2 = \log(2/\delta)(2C_\theta\varepsilon + 2\sigma_\theta^2)$. As a result, we have with confidence $1-\delta$ that

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_\theta(x_i,z_i) - \boldsymbol{\mu}_\theta\right\|_{\mathcal{F}^q} \leq \frac{2C_\theta\log\frac{2}{\delta}}{n} + \sqrt{\frac{2\sigma_\theta^2\log\frac{2}{\delta}}{n}}, \tag{28}$$

as required. □

It remains to show that, for each $\theta \in \Theta$, there exists a constant $C_\theta < \infty$ such that $\|\boldsymbol{\xi}_\theta(X,Z)\|_{\mathcal{F}^q} < C_\theta$ almost surely. Note that for any $(x,z) \in \mathcal{X}\times\mathcal{Z}$ for which $P_{XZ}(x,z) > 0$,

$$\begin{aligned} \|\boldsymbol{\xi}_\theta(x,z)\|_{\mathcal{F}^p} &= \sqrt{\|\boldsymbol{\xi}_\theta(x,z)\|_{\mathcal{F}^p}^2} \\ &= \sqrt{\boldsymbol{\psi}(z;\theta)^\top\boldsymbol{\psi}(z;\theta)k(x,x)} \\ &\leq \sup_{x,z}\sqrt{\boldsymbol{\psi}(z;\theta)^\top\boldsymbol{\psi}(z;\theta)k(x,x)} < \infty, \end{aligned}$$

where the last inequality follows from Assumptions **(A2)** and **(A4)**.

D.3 Proof of Theorem 3.2

Proof. If $\mathscr{M}(x;\theta_1) = \mathscr{M}(x;\theta_2)$ for P_X-almost all x, then the equality $\boldsymbol{\mu}_{\theta_1} = \boldsymbol{\mu}_{\theta_2}$ follows straightforwardly. Suppose that $\boldsymbol{\mu}_{\theta_1} = \boldsymbol{\mu}_{\theta_2}$ and let $\boldsymbol{\delta}(x) := \mathscr{M}(x;\theta_1) - \mathscr{M}(x;\theta_2)$. Then, we have

$$\begin{aligned} \|\boldsymbol{\mu}_{\theta_1} - \boldsymbol{\mu}_{\theta_2}\|_{\mathcal{F}^q}^2 &= \left\|\int \boldsymbol{\xi}_{\theta_1}(x,z)\,\mathrm{d}P_{XZ}(x,z) - \int \boldsymbol{\xi}_{\theta_2}(x,z)\,\mathrm{d}P_{XZ}(x,z)\right\|_{\mathcal{F}^q}^2 \\ &= \left\|\int \mathscr{M}(x;\theta_1)k(x,\cdot)\,\mathrm{d}P_X(x) - \int \mathscr{M}(x;\theta_2)k(x,\cdot)\,\mathrm{d}P_X(x)\right\|_{\mathcal{F}^q}^2 \\ &= \left\|\int (\mathscr{M}(x;\theta_1) - \mathscr{M}(x;\theta_2))k(x,\cdot)\,\mathrm{d}P_X(x)\right\|_{\mathcal{F}^q}^2 \\ &= \iint \boldsymbol{\delta}(x)^\top k(x,x')\boldsymbol{\delta}(x')\,\mathrm{d}P_X(x)\,\mathrm{d}P_{X'}(x') = 0, \end{aligned} \tag{29}$$

where X' is an independent copy of X. It follows from (29) and Assumption **(A2)** that the function $g(x) := \boldsymbol{\delta}(x)p_X(x)$ has zero L2-norm, *i.e.*, $\|g\|_2^2 = 0$ where p_X denotes the density of P_X. As a result, $\boldsymbol{\delta}(x) = \mathbf{0}$ a.e. P_X implying that $P_X(B_0) = 1$ where $B_0 := \{x \in \mathcal{X} : \mathscr{M}(x;\theta_1) - \mathscr{M}(x;\theta_2) = \mathbf{0}\}$. Therefore, $\mathscr{M}(x;\theta_1) = \mathscr{M}(x;\theta_2)$ for P_X-almost all x. This completes the proof. □

D.4 Proof of Theorem 3.3

Proof. By the definition of $\mathbb{M}(\theta)$ and the Bochner integrability of $\boldsymbol{\xi}_\theta$,

$$
\begin{aligned}
\mathbb{M}^2(\theta) &= \|\boldsymbol{\mu}_\theta\|^2_{\mathcal{F}^q} \\
&= \langle \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\theta \rangle_{\mathcal{F}^q} \\
&= \langle \mathbb{E}[\boldsymbol{\xi}_\theta(X,Z)], \mathbb{E}[\boldsymbol{\xi}_\theta(X,Z)] \rangle_{\mathcal{F}^q} \\
&= \mathbb{E}[\langle \boldsymbol{\xi}_\theta(X,Z), \mathbb{E}[\boldsymbol{\xi}_\theta(X,Z)] \rangle_{\mathcal{F}^q}] \\
&= \mathbb{E}[\langle \boldsymbol{\xi}_\theta(X,Z), \boldsymbol{\xi}_\theta(X',Z') \rangle_{\mathcal{F}^q}] \\
&= \mathbb{E}[h_\theta((X,Z),(X',Z'))],
\end{aligned}
$$

where (X',Z') is an independent copy of (X,Z) with an identical distribution. □

D.5 Proof of Theorem 3.4

Proof. By Mercer's theorem [49, Theorem 4.49], we have $k(x,x') = \sum_j \lambda_j e_j(x)e_j(x')$ where the convergence is absolute and uniform. Recall that $\boldsymbol{\zeta}^j_\theta(x,z) := (\psi_1(z;\theta)e_j(x), \ldots, \psi_q(z;\theta)e_j(x))$. Hence, we can express the kernel h_θ as

$$
\begin{aligned}
h_\theta((x,z),(x',z')) &= \boldsymbol{\psi}(z;\theta)^\top \boldsymbol{\psi}(z';\theta) k(x,x') \\
&= \boldsymbol{\psi}(z;\theta)^\top \boldsymbol{\psi}(z';\theta) \left(\sum_j \lambda_j e_j(x) e_j(x') \right) \\
&= \sum_j \lambda_j \boldsymbol{\psi}(z;\theta)^\top \boldsymbol{\psi}(z';\theta) e_j(x) e_j(x') \\
&= \sum_j \lambda_j \left[\boldsymbol{\psi}(z;\theta) e_j(x)\right]^\top \left[\boldsymbol{\psi}(z';\theta) e_j(x')\right] \\
&= \sum_j \lambda_j \boldsymbol{\zeta}^j_\theta(x,z)^\top \boldsymbol{\zeta}^j_\theta(x',z').
\end{aligned}
$$

Since $\lambda_j > 0$, the function h_θ is positive definite. Then, we can express $\mathbb{M}^2(\theta)$ as follows:

$$
\begin{aligned}
\mathbb{M}^2(\theta) &= \mathbb{E}[h_\theta((X,Z),(X',Z'))] \\
&= \mathbb{E}\left[\sum_j \lambda_j \boldsymbol{\zeta}^j_\theta(X,Z)^\top \boldsymbol{\zeta}^j_\theta(X',Z') \right] \\
&= \sum_j \lambda_j \mathbb{E}_{XZ}\left[\boldsymbol{\zeta}^j_\theta(X,Z)\right]^\top \mathbb{E}_{X'Z'}\left[\boldsymbol{\zeta}^j_\theta(X',Z')\right] \\
&= \sum_j \lambda_j \left\| \mathbb{E}_{XZ}\left[\boldsymbol{\zeta}^j_\theta(X,Z)\right] \right\|_2^2 .
\end{aligned}
$$

This completes the proof. □

D.6 Proof of Theorem B.1

In order to show the consistency of $\hat{\theta}_n := \arg\min_{\theta\in\Theta} \widehat{\mathbb{M}}^2_n(\theta)$, we need the uniform consistency of $\widehat{\mathbb{M}}^2_n(\theta)$ and the continuity of $\theta \mapsto \mathbb{M}^2(\theta)$. The following lemma gives these two results.

Lemma D.1. *Assume that there exists an integrable and symmetric function $F_{\boldsymbol{\psi}}$ such that $\|\boldsymbol{\psi}(z,\theta)\|_2 \le F_{\boldsymbol{\psi}}(z)$ for any $\theta \in \Theta$ and $z \in \mathcal{Z}$. If Assumption* (A4) *holds,* $\sup_{\theta\in\Theta} |\mathbb{M}^2_n(\theta) - \mathbb{M}^2(\theta)| \xrightarrow{p} 0$ *and $\theta \mapsto \mathbb{M}^2(\theta)$ are continuous.*

Proof. Recall that

$$\begin{aligned}\mathbb{M}^2(\theta) &= \mathbb{E}[h_\theta((X,Z),(X',Z'))] \\ \widehat{\mathbb{M}}^2_n(\theta) &= \frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n} h_\theta((x_i,z_i),(x_j,z_j)),\end{aligned}$$

where $h_\theta((x,z),(x',z')) = \langle \boldsymbol{\xi}_\theta(x,z), \boldsymbol{\xi}_\theta(x',z')\rangle_{\mathcal{F}^q} = \boldsymbol{\psi}(z;\theta)^\top\boldsymbol{\psi}(z';\theta)k(x,x')$. Then, it follows that

$$\begin{aligned}|h_\theta((x,z),(x',z'))| &= |\langle \boldsymbol{\xi}_\theta(x,z), \boldsymbol{\xi}_\theta(x'z')\rangle_{\mathcal{F}^q}| \\ &\leq \|\boldsymbol{\xi}_\theta(x,z)\|_{\mathcal{F}^q}\cdot\|\boldsymbol{\xi}_\theta(x',z')\|_{\mathcal{F}^q} \\ &= \sqrt{\boldsymbol{\psi}(z;\theta)^\top\boldsymbol{\psi}(z;\theta)k(x,x)}\sqrt{\boldsymbol{\psi}(z';\theta)^\top\boldsymbol{\psi}(z';\theta)k(x',x')} \\ &= \|\boldsymbol{\psi}(z;\theta)\|_2\|\boldsymbol{\psi}(z';\theta)\|_2\sqrt{k(x,x)k(x',x')} \\ &\leq F_{\boldsymbol{\psi}}(z)F_{\boldsymbol{\psi}}(z')\sqrt{k(x,x)k(x',x')},\end{aligned}$$

where $F_{\boldsymbol{\psi}}$ is an integrable and symmetric function. By Assumption **(A4)**, $(x,x')\mapsto\sqrt{k(x,x)k(x',x')}$ is also an integrable function. Hence, h_θ is integrable. Since Θ is compact, it then follows from Newey and McFadden [39, Lemma 2.4] that $\sup_{\theta\in\Theta}|\widehat{\mathbb{M}}^2_n(\theta)-\mathbb{M}^2(\theta)|\xrightarrow{\text{p}} 0$ and $\theta\mapsto\mathbb{M}^2(\theta)$ is continuous. □

Now, we are in the position to present the proof of Theorem B.1.

Proof of Theorem B.1. By Assumption **(A3)** and Theorem 3.2, $\mathbb{M}^2(\theta)=0$ if and only if $\theta=\theta_0$. Thus $\mathbb{M}^2(\theta)$ is uniquely minimized at θ_0. Since Θ is compact, $\mathbb{M}^2(\theta)$ is continuous and $\widehat{\mathbb{M}}^2_n(\theta)$ converges uniformly in probability to $\mathbb{M}^2(\theta)$ by Lemma D.1. Then, $\hat{\theta}_n\xrightarrow{\text{p}}\theta_0$ by Newey and McFadden [39, Theorem 2.1]. □

D.7 Proof of Theorem 4.1

Proof. First, we need to check that $\sigma_h^2\neq 0$ when $\theta\neq\theta_0$ and $\sigma_h^2=0$ when $\theta=\theta_0$. Then, the results follow directly from Serfling [44, Sec. 5.5.1 and Sec. 5.5.2].

Note that $\mathbb{E}_{u'}[h_\theta(u,u')] = \mathbb{E}_{u'}[\langle\boldsymbol{\xi}_\theta(u),\boldsymbol{\xi}_\theta(u')\rangle_{\mathcal{F}^q}] = \langle\boldsymbol{\xi}_\theta(u),\mathbb{E}_{u'}[\boldsymbol{\xi}_\theta(u')]\rangle_{\mathcal{F}^q} = \langle\boldsymbol{\xi}_\theta(u),\boldsymbol{\mu}_\theta\rangle_{\mathcal{F}^q} = M_\theta\boldsymbol{\xi}_\theta(u)$. When $\theta=\theta_0$, it follows that $\mathbb{E}_{u'}[h_{\theta_0}(u,u')]=0$ by Lemma 3.1, and hence $\sigma_h^2=0$.

Next, suppose that $\theta\neq\theta_0$. Then, $\mathbb{E}_{u'}[h_\theta(u,u')] = M_\theta\boldsymbol{\xi}_\theta(u) =: c(u)$. Since $\sigma_h^2 = \mathrm{Var}_u[c(u)] = \mathbb{E}_u[(c(u)-\mathbb{E}_{u'}[c(u')])^2]$, $\sigma_h^2=0$ if and only if $c(u)$ is a constant function. Note that we can write $c(u)=c(x,z)=\mathbb{E}_{X'Z'}[\boldsymbol{\psi}(Z';\theta)^\top\boldsymbol{\psi}(z;\theta)k(x,X')]$. Therefore, by Assumptions **(A3)** and **(A4)**, $c(u)$ cannot be a constant function, implying that $\sigma_h^2>0$. □

D.8 Proof of Theorem 5.1

Proof. Since the kernel $k(x,x')=\varphi(x-x')$ is a shift-invariant kernel on $\mathbb{R}^d$, it follows from Theorem 2.1 that

$$\varphi(x-x') = (2\pi)^{-d/2}\int_{\mathbb{R}^d} e^{-i(x-x')^\top\omega}\,\mathrm{d}\Lambda(\omega).$$

Therefore, we can express $\mathbb{M}^2(\theta)$ as

$$\begin{aligned}\mathbb{M}^2(\theta) &= \mathbb{E}[\boldsymbol{\psi}(Z;\theta)^\top\boldsymbol{\psi}(Z';\theta)k(X,X')] \\ &= \mathbb{E}[\boldsymbol{\psi}(Z;\theta)^\top\boldsymbol{\psi}(Z';\theta)\varphi(X-X')] \\ &= (2\pi)^{-d/2}\mathbb{E}\left[\boldsymbol{\psi}(Z;\theta)^\top\boldsymbol{\psi}(Z';\theta)\left(\int_{\mathbb{R}^d} e^{-i(X-X')^\top\omega}\,\mathrm{d}\Lambda(\omega)\right)\right] \\ &= (2\pi)^{-d/2}\mathbb{E}\left[\boldsymbol{\psi}(Z;\theta)^\top\boldsymbol{\psi}(Z';\theta)\left(\int_{\mathbb{R}^d} e^{-i\omega^\top X}\cdot e^{i\omega^\top X'}\,\mathrm{d}\Lambda(\omega)\right)\right]\end{aligned}$$

$$
\begin{aligned}
&= (2\pi)^{-d/2}\mathbb{E}\left[\int_{\mathbb{R}^d} \boldsymbol{\psi}(Z;\theta)^\top \boldsymbol{\psi}(Z';\theta) e^{-i\omega^\top X} e^{i\omega^\top X'} \,\mathrm{d}\Lambda(\omega)\right] \\
&= (2\pi)^{-d/2}\mathbb{E}\left[\int_{\mathbb{R}^d} \left[\boldsymbol{\psi}(Z;\theta) e^{-i\omega^\top X}\right]^\top \left[\boldsymbol{\psi}(Z';\theta) e^{i\omega^\top X'}\right] \,\mathrm{d}\Lambda(\omega)\right] \\
&= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathbb{E}\left[\boldsymbol{\psi}(Z;\theta) e^{-i\omega^\top X}\right]^\top \mathbb{E}\left[\boldsymbol{\psi}(Z';\theta) e^{i\omega^\top X'}\right] \,\mathrm{d}\Lambda(\omega) \\
&= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \left\|\mathbb{E}\left[\boldsymbol{\psi}(Z;\theta)\exp(i\omega^\top X)\right]\right\|_2^2 \,\mathrm{d}\Lambda(\omega).
\end{aligned}
$$

This completes the proof. □