Electronic Journal of Statistics
Vol. 16 (2022) 5277–5320
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2058

Testing subspace restrictions in the presence of high dimensional nuisance parameters

Alessio Sancetta

Department of Economics, Royal Holloway University of London, Egham TW20 0EX, UK
e-mail: asancetta@gmail.com

Abstract: A framework for hypothesis testing of functional restrictions against general alternatives is proposed. The parameter space is a subset of a reproducing kernel Hilbert space (RKHS). The null hypothesis does not necessarily define a parametric model. The test allows us to deal with possibly infinite dimensional nuisance parameters. The methodology is based on a moment equation similar in spirit to the construction of the efficient score in semiparametric statistics. The feasible version of such moment equation requires to consistently estimate projections in the space of RKHS. A tractable asymptotic theory is established for this problem. Simulation results show that the finite sample performance of the test is consistent with the asymptotic results and that ignoring the effect of nuisance parameters highly distorts the size of the tests.

MSC2020 subject classifications: Primary 62G10; secondary 62F03.
Keywords and phrases: Functional restriction, hypothesis testing, projection operator, nonlinear model, reproducing kernel Hilbert space.

Received September 2021.

Contents

1. Introduction

Suppose that we are interested in estimating the number of event arrivals Y over some interval, conditioning on a vector of covariates X known at the start of the interval. We decide to minimize the negative log-likelihood for Poisson arrivals with conditional intensity $\exp\{\mu(X)\}$ for some function μ. For observation i, the negative loglikelihood is proportional to

$$\exp\{\mu(X_i)\} - Y_i\mu(X_i). \tag{1.1}$$

We suppose that μ lies in some infinite dimensional space. For example, to avoid the curse of dimensionality, we could choose

$$\mu(X) := \sum_{k=1}^{K} f^{(k)}\left(X^{(k)}\right) \tag{1.2}$$

where $X^{(k)}$ denotes the k^{th} covariate (the k^{th} element of the K-dimensional covariate X), and the univariate functions $f^{(k)}$ are elements in some possibly infinite dimensional space. However, we restrict $f^{(1)}$ to be a linear function. Then, we want to test whether linearity with respect to the first variable holds against the alternative of a general additive model. We could also test against the alternative of a general continuous multivariate function, not necessarily

additive. This paper addresses practical problems such as the above. The paper is not restricted to this Poisson problem or additive models on real valued variables.

From the example above, we need to (i) estimate μ, which in this example we chose to be additive with $f^{(1)}$ linear under the null; we need to (ii) test this additive restriction, against a more general non-parametric alternative. Under the null, the remaining $K-1$ functions in (1.2) are not specified. Problem (i) is standard. Having solved problem (i), the solution to problem (ii) requires to test a non-parametric hypothesis (an additive model with linear $f^{(1)}$) with infinite dimensional nuisance parameters (the remaining unknown $K-1$ functions) against a more general non-parametric alternative. In this paper, we shall call the restriction under the null semiparametric. This does not necessarily mean that the parameter of interest is finite dimensional, as often the case in the semiparametric literature.

Semiparametric inference requires that the infinite dimensional parameter and the finite dimensional one are orthogonal in the population [2, Equation (2.12)]). In our Poisson motivating example this is not the case. Even if the restriction is parametric, we do not need to suppose that the parameter value is known under the null. This requires us to modify the test statistic in order to achieve the required orthogonality. Essentially, we project the test statistic on some space that is orthogonal to the possibly infinite dimensional nuisance parameter. This is the procedure involved in the construction of the efficient score in semiparametric statistics. The reader is referred to [30] for a review of the basic idea. Here, we are concerned with functional restrictions and are able to obtain critical values by fast simulation.

Under the null hypothesis, we can find a representation for the limiting asymptotic distribution which is amenable of fast simulation. In consequence critical values do not need to be generated using resampling procedures. While the discussion of the asymptotic validity of the procedure is involved, the implementation of the test is simple. The Matlab code to perform the test and compute its critical values is available from the https://github.com/asancetta/ARKHS/. A set of simulations confirm that the procedure works well, and illustrates the well known fact that nuisance parameters can considerably distort the size of a test if not accounted for using our proposed procedure.

1.1. Relation to the literature

The test statistic of this paper is in the form of a restricted score test for high possibly infinite dimensional parameter space under the null and alternative in order to test restrictions. The approach used here is inspired by the definition of efficient score in semiparametric estimation. In the case of parametric inference, the present approach includes the projection method described in [34]. Here, we consider more general problems for possibly infinite dimensional estimators beyond nonlinear regression.

The problem of testing the restricted functional form of an unknown function against more general alternatives is an extensively studied problem (e.g. [5], [34],

[16], [27], [28], [29], [11], [12]). A common approach is to generate some form of residuals that are supposed to be orthogonal to test functions that define the space of alternatives. This is tantamount to the construction of a score test or some function of it. With different level of generality the above references consider purely parametric approaches, but some allow for nonparametric alternatives. General specifications under the null are also considered in [11] and [12]. Some of these approaches exploit the properties of kernel smoothers and the fact that the parameter to be estimated is a conditional expectation of a response variable. Moreover, they also tend to rely on the bootstrap to derive p-values.

General results in the context of high dimensional models can be found in [4]. There, the reader can also find the main references in that literature. The asymptotic distribution requires the use of the bootstrap in order to compute critical values.

Inspired by the early statistical contributions of Neyman, a number of authors have recently used the term Neyman orthogonality to refer to the independence between parameters of interest and nuisance parameters. For the estimation problem in the presence of a functional nuisance parameter, [8] propose a correction to achieve Neyman orthogonality using sample splitting.

In [13] a generalized Likelihood Ratio test of the null of parametric or nonparameteric additive restrictions versus general nonparametric ones is developped. This is based on a Gaussian error model (or parametric error distribution) for additive regression, and estimation using smoothing kernels. This approach has been extended to the nonparametric error distribution in [14]. The asymptotic distribution is Chi-square with degrees of freedom equal to some (computable) function of the the data. In [7] the framework of sieve estimation is considered and a likelihood ratio statistic with asymptotic Chi-square distribution is derived (see also [25]).

The approach considered here is complementary to the above references. It allows the parameter space to be a RKHS of smooth functions. These include functions in Hilbert Sobolev spaces. Estimation in RKHS is well understood and can cater for many circumstances of interest in applied work ([33], [26]). For example, it is possible to view sieve estimation as estimation in RKHS where the feature space defined by the kernel increases with the sample size. The testing procedure is based on a corrected moment condition. Hence, it does not rely on likelihood estimation. The conditions used are elementary, as they just require existence of real valued derivatives of the loss function (in the same vein as [9]) and mild regularity conditions on the covariance kernel. The correction is estimated by either ridge regression, or just ordinary least square using pseudo-inverse. The estimation and testing procedure makes use of the full sample with no need to use either the bootstrap or sample splitting. However, it does require to simulate from a linear combination of i.i.d. chi-square random variables.

The problem of testing relying on the machinery of RKHS is also discussed in [18]. There, interest lies in constructing confidence regions on the true parameter based on a score test statistic. As far as overlap is concerned, we note that their theoretical results (Theorems 1 and 2) apply to the case when the true

parameter is known under the null. In practice to make the test operational, one needs to replace the true value with an estimator. To test restrictions, we need to account for the estimation of the true unknown parameter under the null. Then, a correction is usually required. Regarding the distribution of the test statistic, the results in [18] rely on the use of the bootstrap. On the other hand, here, we derive the exact limiting distribution of our test statistic. This can then be simulated without the need to estimate the model multiple times as with the bootstrap. Given the difference in focus, it is difficult to make a comparison beyond the aforementioned remarks.

1.2. Outline

The plan for the paper is as follows. Section 2 discusses the testing problem. Section 2.1 provides a motivating simulation example to show that even in simple problems, the nuisance parameter can distort the asymptotic distribution of a test. Section 2.2 reviews some basic facts of RKHS. The remaining of Section 2 defines the problem, the test, and describes its implementation. Section 3 contains the asymptotic analysis of the proposed testing procedure. Section 4 concludes with a finite sample analysis via simulations. Section 6 concludes the paper. The proofs, and additional results are in the Appendix.

2. The inference problem

The explanatory variable $X^{(k)}$ takes values in $\mathcal{X}$, a compact subset of a separable Banach space $(k = 1, 2, ..., K)$. The most basic example of $\mathcal{X}$ is $[0, 1]$. The vector covariate $X = \left(X^{(1)}, ..., X^{(K)}\right)$ takes values in the Cartesian product $\mathcal{X}^K$, e.g., $[0, 1]^K$. The dependent variable takes values in $\mathcal{Y}$, usually a subset of $\mathbb{R}$. Let $Z := (Y, X)$. This takes values in $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}^K$. If no dependent variable Y can be defined (e.g., unsupervised learning, or certain likelihood estimators), $Z = X$. Let P be the law of Z, and use linear functional notation, i.e., for any $f : \mathcal{Z} \to \mathbb{R}$, $Pf = \int_{\mathcal{Z}} f(z)\, dP(z)$. Let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$, where δ_{Z_i} is the point mass at Z_i, implying that $P_n f = \frac{1}{n}\sum_{i=1}^n f(Z_i)$ is the sample mean of $f(Z)$. For $p \in [1, \infty]$, let $|\cdot|_p$ be the L_p norm (w.r.t. the measure P), e.g., for $f : \mathcal{Z} \to \mathbb{R}$, $|f|_p = \left(P|f|^p\right)^{1/p}$, with the obvious modification to sup norm when $p = \infty$. We shall also abbreviate left hand side, right hand side and with respect to with l.h.s., r.h.s., and w.r.t., respectively.

2.1. Motivation

The problem can be described as follows, though in practice we will need to add extra regularity conditions. Let $\mathcal{H}^K$ be a vector space of real valued functions on $\mathcal{X}^K$, equipped with a norm $|\cdot|_{\mathcal{H}^K}$. Consider a loss function $L : \mathcal{Z} \times \mathbb{R} \to \mathbb{R}$. We shall be interested in the case where the second argument is $\mu(x)$: $L(z, \mu(x))$ with $\mu \in \mathcal{H}^K$. Therefore, to keep notation compact, let $\ell_\mu(Z) = L(Z, \mu(X))$.

For the special case of the square error loss we would have $\ell_\mu(z) = L(z, \mu(x)) = |y - \mu(x)|^2$ $(z = (y, x))$. The use of ℓ_μ makes it more natural to use linear functional notation. The unknown function of interest is the minimizer μ_0 of $P\ell_\mu$, and it is assumed to be in $\mathcal{H}^K$. Let $\mathcal{H}^K(B) := \left\{\mu \in \mathcal{H}^K : |\mu|_{\mathcal{H}^K} \leq B\right\}$ be the ball of $|\cdot|_{\mathcal{H}^K}$-radius B in $\mathcal{H}^K$. Then,

$$\mu_0 = \arg \inf_{\mu \in \mathcal{H}^K(B)} P\ell_\mu. \tag{2.1}$$

This minimizer always exists and is unique under regularity conditions on the loss because $\mathcal{H}^K(B)$ is closed. The main goal it to test the restriction that $\mu \in \mathcal{R}_0(B) := \mathcal{R}_0 \cap \mathcal{H}^K(B)$ where $\mathcal{R}_0$ is some subspace of $\mathcal{H}^K$ (for example a linear restriction).

Let $\partial^k \ell_\mu(z) = \partial^k L(z,t)/\partial t^k\big|_{t=\mu(x)}$ be the k^{th} partial derivative of $L(z,t)$ with respect to t and then evaluated at $\mu(x)$. The validity of this derivative and other related quantities will be ensured by the regularity conditions that we shall impose. By the first order conditions, the optimizer in (2.1) satisfies the equality $P\partial\ell_{\mu_0} h = 0$ for any $h \in \mathcal{H}^K$ when μ_0 is in the interior of $\mathcal{H}^K(B)$, which we shall write as $\mu_0 \in \operatorname{int}\left(\mathcal{H}^K(B)\right)$. This is how the population version of Z-estimators in Banach spaces is defined ([31, Chapter 3.3]. When $\mu_0 \in \operatorname{int}(\mathcal{R}_0(B))$, (2.1) is the same as $\mu_0 = \arg\inf_{\mu \in \mathcal{R}_0} P\ell_\mu$. Given that the law P is unknown, we rely on the sample data to test if the restriction holds.

To this end, we find an estimator $\mu_{0n} = \arg\inf_{\mu \in \mathcal{R}_0(B)} P_n \ell_\mu$. To test the restriction we can look at how close

$$\sqrt{n} P_n \partial\ell_{\mu_{0n}} h = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial\ell_{\mu_{0n}}(Z_i)\, h(X_i) \tag{2.2}$$

is to zero for suitable choice of $h \in \mathcal{H}^K \setminus \mathcal{R}_0$. By the aforementioned remarks, we also have that μ_{0n} satisfies $P_n \partial\ell_{\mu_{0n}} h = 0$ for $h \in \mathcal{R}_0$ if $\mu_{0n} \in \operatorname{int}(\mathcal{R}_0(B))$. Hence we may restrict attention to $h \in \mathcal{H}^K \setminus \mathcal{R}_0$. The compact notation on the l.h.s. of (2.2) shall be used throughout the paper.

A test statistic can be constructed from (2.2) as follows:

$$\frac{1}{R} \sum_{r=1}^{R} \left(\sqrt{n} P_n \partial\ell_{\mu_{0n}} h^{(r)}\right)^2 \tag{2.3}$$

where $h^{(r)} \in \mathcal{H}^K \setminus \mathcal{R}_0$, $r = 1, 2, ..., R$ and R is an arbitrary integer, possibly much larger than n. The covariance matrix of the process in (2.2) is singular, when R is large and the space of test functions is compact in a suitable topology. For this reason, it is necessary to use the Cramer-von Mises' type of statistic in (2.3) instead of quadratic form based on the inverse of the covariance matrix of the score statistic. We discuss alternative statistics, including the use of a pivotal statistic, in the paragraph Additional Remarks in Section 3.2.

If μ_{0n} is orthogonal to the functions $h \in \mathcal{H}^K \setminus \mathcal{R}_0$ ([2, Equation 2.12]), (2.2) is, to first order, equal in distribution to $\sqrt{n} P_n \partial\ell_{\mu_0} h$. Hence, it is simple to find

the asymptotic distribution of (2.2) and hence of (2.3). Supposing asymptotic stochastic equicontinuity and the null that $\sqrt{n}P\partial\ell_{\mu_0}h=0$, it can be shown that ([31, Theorem 3.3.1]),

$$\sqrt{n}P_n\partial\ell_{\mu_{0n}}h=\sqrt{n}P_n\partial\ell_{\mu_0}h+\sqrt{n}P\partial^2\ell_{\mu_0}\left(\mu_{0n}-\mu_0\right)h+o_p\left(1\right). \tag{2.4}$$

If the second term on the r.h.s. were zero, the limiting distribution of $\sqrt{n}P_n\partial\ell_{\mu_{0n}}h$ would be the same of the one of $\sqrt{n}P_n\partial\ell_{\mu_0}h$, which does not depend on the nuisance parameter μ_{0n}. For example, the orthogonality condition in [2, Equation 2.12] guarantees that the second term on the r.h.s. is zero ([2, Equation 2.8], assuming Fréchet differentiability). Such orthogonality condition has been referred to as Neyman orthogonality in recent literature that uses machine learning-based estimators [8].

Such condition does not always hold, implying that the asymptotic distribution is more complex as it requires to account for the extra term $\sqrt{n}P\partial^2\ell_{\mu_0}\left(\mu_{0n}-\mu_0\right)h$. Not accounting for such term, as if there were no nuisance parameter, can seriously distort the distribution of the test. However, the distribution of the sum of the two terms on the r.h.s. of (2.4) cannot easily be found in general. In the special case of the nonparametric regression with a higher order kernel, [12] shows that (2.4) converges to $\sqrt{n}P_n\partial\ell_{\mu_0}\tilde{h}$ where $\tilde{h}$ is a linear functional of h. Their argument is specific to their setup and does not extend to the present framework. To address the problem in a simple and general way, we remove the effect of the nuisance parameter constructing functions $h\in\mathcal{H}^K\setminus\mathcal{R}_0$ such that the second term on the r.h.s. of (2.4) is zero in a way that does not affect the power of the test under the alternative that the restriction does not hold. This is tantamount to finding functions $h\in\mathcal{H}^K\setminus\mathcal{R}_0$ that are orthogonal to functions in $\mathcal{R}_0$ in the sense that that they asymptotically satisfy

$$P\partial^2\ell_{\mu_0}\nu h=0 \tag{2.5}$$

for any $\nu\in\mathcal{R}_0$ when $\mu_0\in\operatorname{int}\left(\mathcal{R}_0\left(B\right)\right)$. The challenge is that the set of such orthogonal functions $h\in\mathcal{H}^K\setminus\mathcal{R}_0$ needs to be estimated. It is not clear before hand that estimation of this set of functions leads to the same asymptotic distribution as if this set were known. We show that this is the case. We define a sample based linear operator $\Pi_{n,\rho}:\mathcal{H}^K\to\mathcal{R}_0$ such that

$$\begin{aligned}\Pi_{n,\rho}h&=\arg\inf_{\nu\in\mathcal{R}_0}P_n\partial^2\ell_{\mu_{0n}}\left(h-\nu\right)^2+\rho\left|\nu\right|_{\mathcal{H}^K}^2 \\ &=:\arg\inf_{\nu\in\mathcal{R}_0}\frac{1}{n}\sum_{i=1}^{n}\partial^2\ell_{\mu_{0n}}\left(Z_i\right)\left(h\left(X_i\right)-\nu\left(X_i\right)\right)^2+\rho\left|\nu\right|_{\mathcal{H}^K}^2,\end{aligned} \tag{2.6}$$

which depends on $\rho=\rho_n\to 0$. The suitable rate is given in Theorem 2 and Corollary 1. Given a set of functions $\left\{h^{(r)}\in\mathcal{H}^K\setminus\mathcal{R}_0:r=1,2,...,R\right\}$ we use (2.6) to define $\left\{\hat{h}^{(r)}=h^{(r)}-\Pi_{n,\rho}h^{(r)}:r=1,2,...,R\right\}$. We show that this is a set of asymptotically orthogonal functions in the sense of (2.5) for any $h=\hat{h}^{(r)}$, $r=1,2,...,R$ when $n\to\infty$.

Then, we suggest to replace the test statistic (2.3) with

$$\hat{S}_n = \frac{1}{R}\sum_{r=1}^{R}\left(\sqrt{n}P_n\partial\ell_{\mu_{0n}}\hat{h}^{(r)}\right)^2. \tag{2.7}$$

We give the asymptotic distribution of (2.7) and show that this distribution can be easily simulated.

The null hypothesis $\mu_0 \in \mathcal{R}_0$ is equivalent to $P\partial\ell_{\mu_0}h = 0$ for $h \in \mathcal{H}^K \setminus \mathcal{R}_0$. Hence, deviations should only be detected for such test functions h. In (3.1) in Section 3.2, we define the population version of (2.6), which we denote by Π_0. For any $h \in \mathcal{H}^K$ we have that $h = (h - \Pi_0 h) + \Pi_0 h$, where by definition $\Pi_0 h \in \mathcal{R}_0$. Hence, the only deviations from the null hypotheses that matter are in the direction of $(h - \Pi_0 h)$. This is exactly what the test statistic in (2.7) attempts to do.

In summary, given functions $h^{(r)} \in \mathcal{H}^K$, the statistic (2.3) uses $h = h^{(r)}$, while (2.7) uses $h = h^{(r)} - \Pi_{n,\rho}h^{(r)}$, $r = 1, 2, ..., R$. The latter is not affected by the second term in (2.4) so that we can use the asymptotic distribution of $\sqrt{n}P_n\partial\ell_{\mu_0}h$ to find the p-values

In Section 4 we provide some finite sample evidence to show that using the asymptotic distribution of $\sqrt{n}P_n\partial\ell_{\mu_0}h$ to find p-values leads to distortions in size when the test functions h are not adjusted for the presence of nuisance parameters. This is the case even for simple problems.

2.2. *Additional notation and basic facts about reproducing kernel Hilbert spaces*

This paper uses RKHS in order to control the complexity of the estimators in a general way. Intuitively, we can think of RKHS as Hilbert spaces with additional smoothness constraints. RKHS provide an explicit representation for functions in common function spaces, such as Hilbert Sobolev spaces. Estimators that are minimizers within many function spaces are elements in RKHS and can possibly have an explicit solution. Rather than focusing on some special function spaces such as Hilbert Sobolev spaces, this paper considers these more general spaces. Moreover, estimators such as splines are just RKHS estimators ([33]). Next, we discuss some basic facts about RKHS. The discussion should clarify the aforementioned remarks.

A RKHS of bounded functions is uniquely generated by a centered Gaussian measure with covariance C [20] and C is usually called the (reproducing) kernel of $\mathcal{H}$. We consider covariance functions with representation

$$C(s,t) = \sum_{v=1}^{\infty}\lambda_v^2\varphi_v(s)\varphi_v(t), \tag{2.8}$$

for linearly independent functions $\varphi_v : \mathcal{X} \to \mathbb{R}$ and coefficients λ_v such that $\sum_{v=1}^{\infty}\lambda_v^2\varphi_v^2(s) < \infty$. Here, linear independent means that if there is a sequence

of real numbers $(f_v)_{v\geq 1}$ such that $\sum_{v=1}^{\infty} f_v^2/\lambda_v^2 < \infty$ and $\sum_{v=1}^{\infty} f_v \varphi_v(s) = 0$ for all $s \in \mathcal{X}$, then $f_v = 0$ for all $v \geq 1$. The coefficients λ_v^2 would be the eigenvalues of (2.8) if the functions φ_v were orthonormal, but this is not implied by the above definition of linear independence. The restriction to (2.8) only means that the function space is separable. The RKHS $\mathcal{H}$ is the completion of the set of functions representable as $f(x) = \sum_{v=1}^{\infty} f_v \varphi_v(x)$ for real valued coefficient f_v such that $\sum_{v=1}^{\infty} f_v^2/\lambda_v^2 < \infty$. Equivalently, $f(x) = \sum_{j=1}^{\infty} \alpha_j C(s_j, x)$, for coefficients s_j in $\mathcal{X}$ and real valued coefficients α_j satisfying $\sum_{j=1}^{\infty} \alpha_i \alpha_j C(s_i, s_j) < \infty$. In fact, for C in (2.8),

$$\sum_{j=1}^{\infty} \alpha_j C(s_j, x) = \sum_{v=1}^{\infty} \left(\sum_{j=1}^{\infty} \alpha_j \lambda_v^2 \varphi_v(s_j) \right) \varphi_v(x) = \sum_{v=1}^{\infty} f_v \varphi_v(x) \tag{2.9}$$

by obvious definition of the coefficients f_v. The change of summation is possible by the aforementioned restrictions on the coefficients λ_v and functions φ_v. The inner product in $\mathcal{H}$ is denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and satisfies $f(x) = \langle f, C(x, \cdot) \rangle_{\mathcal{H}}$. This implies the reproducing kernel property $C(s,t) = \langle C(s, \cdot), C(t, \cdot) \rangle_{\mathcal{H}}$. Therefore, the square of the RKHS norm is defined in the two following equivalent ways

$$|f|_{\mathcal{H}}^2 = \sum_{v=1}^{\infty} \frac{f_v^2}{\lambda_v^2} = \sum_{i,j=1}^{\infty} \alpha_i \alpha_j C(s_i, s_j) \tag{2.10}$$

Throughout, the unit ball of $\mathcal{H}$ will be denoted by $\mathcal{H}(1) := \{f \in \mathcal{H} : |f|_{\mathcal{H}} \leq 1\}$.

The additive RKHS is generated by the Gaussian measure with covariance function $C_{\mathcal{H}^K}(s,t) = \sum_{k=1}^{K} C^{(k)}\left(s^{(k)}, t^{(k)}\right)$, where $C^{(k)}\left(s^{(k)}, t^{(k)}\right)$ is a covariance function on $\mathcal{X} \times \mathcal{X}$ (as C in (2.8)) and $s^{(k)}$ is the k^{th} element in $s \in \mathcal{X}^K$. The RKHS of additive functions is denoted by $\mathcal{H}^K$, which is the set of functions as in (1.2) such that $f^{(k)} \in \mathcal{H}$ and $\sum_{k=1}^{K} \left|f^{(k)}\right|_{\mathcal{H}}^2 < \infty$. For such functions, the inner product is $\langle f, g \rangle_{\mathcal{H}^K} = \sum_{k=1}^{K} \left\langle f^{(k)}, g^{(k)} \right\rangle_{\mathcal{H}}$, where – for ease of notation – the individual RKHS are supposed to be the same. However, in some circumstances, it can be necessary to make the distinction between the spaces (see Example 9 in Section 2.6). The norm $|\cdot|_{\mathcal{H}^K}$ on $\mathcal{H}^K$ is the one induced by the inner product.

Within this scenario, the space $\mathcal{H}^K$ restricts functions to be additive, where the additive functions in $\mathcal{H}$ can be multivariate functions.

Example 1. *Suppose that $K = 1$ and $\mathcal{X} = [0,1]^d$ ($d > 1$) (only one additive function, which is multivariate). Let $C(s,t) = \exp\left\{-a \sum_j |s_j - t_j|^2\right\}$ where s_j is the j^{th} element in $s \in [0,1]^d$, and $a > 0$. This RKHS is of special interest, as it is dense in the space of continuous bounded functions on $[0,1]^d$ (e.g. [9]). A (kernel) C with such property is called universal.*

The framework also covers the case of functional data because $\mathcal{X}$ is a compact subset of a Banach space (see for example [6]). Most problems of interest where

the unknown parameter μ is a smooth function are covered by the current scenario.

Many spaces of interest are equivalent to RKHS with kernel in (2.8) that satisfy certain conditions.

Example 2. *Suppose that $\mathcal{H}$ is the Sobolev Hilbert space of index V on $[0,1]$, i.e. functions with $V \geq 1$ square integrable weak derivatives. This is a RKHS with $C(s,t) = \sum_{v=1}^{V-1} (s,t)^v / (v!)^2 + H_V(s,t)$ where H_V is the covariance function of the $(V-1)$-fold integrated Brownian motion. In particular,*

$$H_V(\cdot,\cdot) = \int_0^1 G_V(\cdot,u) G_V(\cdot,u)\, du \text{ with } G_V(r,u) := \max\left\{ \frac{(r-u)^{V-1}}{(V-1)!}, 0 \right\},$$

where $r,u \in [0,1]$ ([33, Pages 7–8]). Then, the covariance C admits an expansion as in (2.8) with $\lambda_v \lesssim v^{-\eta}$ where $\eta = V$ [22, Corollary 2, and Pages 523–524].

Example 2 shows that if we are interested in restrictions in subspaces of Hilbert Sobolev spaces, we can use the results of this paper seamlessly.

Finally, for some estimation problems, we may use an alternative norm to define the constraint in the estimation. We define $\mathcal{L}^K(B) := \left\{ f \in \mathcal{H}^K : |f|_{\mathcal{L}^K} \leq B \right\}$, where $|f|_{\mathcal{L}^K} := \sum_{k=1}^K \left| f^{(k)} \right|_{\mathcal{H}}$. For finite K, $\mathcal{L}^K = \mathcal{H}^K$, however, it is easy to see that $\mathcal{L}^K(B) \subset \mathcal{H}^K(B)$ for any finite B (e.g. [23, Lemma 1]. When the elements in the series expansion in (2.8) are known, there are approximation algorithms for estimation in $\mathcal{L}^K(B)$ (e.g. [23], and references therein). We shall estimate the nuisance parameter constrained to be in $\mathcal{L}^K(B)$ in one of the simulation examples.

2.3. The restricted estimator in RKHS

We consider inference on functional restrictions possibly allowing μ not to be fully specified under the null. Within this framework, tests based on the moment equation $P_n \partial \ell_\mu h$ for suitable test functions h are natural (recall (2.7)). Let $\mathcal{R}_0, \mathcal{R}_1 \subset \mathcal{H}^K$ be RKHS with covariance kernels $C_{\mathcal{R}_0}$ and $C_{\mathcal{R}_1}$ such that we can write $C_{\mathcal{H}^K} = C_{\mathcal{R}_0} + C_{\mathcal{R}_1}$. Under the null hypothesis we suppose that $\mu_0 \in \mathcal{R}_0$ (μ_0 as in (2.1)). Under the alternative, $\mu_0 \notin \mathcal{R}_0$. Define

$$\mu_{0n} := \arg \inf_{\mu \in \mathcal{R}_0(B)} P_n \ell_\mu. \tag{2.11}$$

Recall that $\mathcal{R}_0(B) := \mathcal{R}_0 \cap \mathcal{H}^K(B)$. This is the estimator under the null hypothesis. For this estimation, we use the kernel $C_{\mathcal{R}_0}$. The penalised estimator is defined as

$$\mu_{n,\tau} = \arg \inf_{\mu} \left\{ P_n \ell_\mu + \tau |\mu|^2_{\mathcal{H}^K} \right\} \tag{2.12}$$

By duality, there is a sample dependent Lagrange multiplier $\tau_{B,n}$ such that $\mu_{0n} = \mu_{n,\tau}$ when $\tau = \tau_{B,n}$. Theorem 1 in [24] says that the solution to the penalized problem takes the form $\mu_{n,\tau}(x) = \sum_{i=1}^n \hat{a}_i C_{\mathcal{R}_0}(X_i, x)$ for sample

dependent real valued coefficients $\hat{a}_i$. Hence, even if the parameter space where the estimator lies is infinite dimensional, μ_{n0} is not. This fact allows us to use matrix algebra to implement the testing problem, as shown in Section 2.4.

2.4. Implementation of the test

We show how to construct the statistic in (2.7) using matrix notation. We then provide a few examples on how to estimate the projection for a number of problems. All such problems require to find μ_{0n} and $\partial^2 \ell_\mu$. These can then be plugged in (2.6). We assume that μ_{0n} has been estimated. This can be done in a variety of ways ([23] for some examples and references).

2.4.1. Estimation of the restricted estimator

We show how to construct the statistic in (2.7) using matrix notation. As discussed in Section 2.3, the estimator takes the form $\mu_{0n}(\cdot) = \sum_{i=1}^{n} \hat{a}_i C_{\mathcal{R}_0}(X_i, \cdot)$ and it is in $\mathcal{R}_0(B)$. By the reproducing kernel property, $|\mu_{0n}|^2_{\mathcal{H}} = \hat{\mathbf{a}}^T \mathbf{C}_0 \hat{\mathbf{a}} \leq B^2$, where $\hat{\mathbf{a}}$ is an $n \times 1$ vector with i^{th} entry equal to a_i; $\mathbf{C}_0$ be the $n \times n$ matrix with (i,j) entry equal to $C_{\mathcal{R}_0}(X_i, X_j)$; the superscript ^{T} is used for transposition. Hence, in (2.12), $\rho_{B,n}$ is chosen such that the aforementioned inequality holds.

For the regression problem under the square error loss, the solution vector $\hat{\mathbf{a}}$ has a closed form: $\hat{\mathbf{a}} := (\mathbf{C}_0 + \rho_{B,n}\mathbf{I})^{-1} \mathbf{y}$ where the $n \times 1$ vector $\mathbf{y}$ has i^{th} entry equal to Y_i. In this case, if the constraint $\{\mu \in \mathcal{R}_0(B)\}$ is binding, the $\rho_{B,n}$ in (2.12) that satisfies the constraint is given by the solution of

$$\sum_{i=1}^{n} \left(\mathbf{y}^T \mathbf{Q}_i\right)^2 \frac{\kappa_i}{\kappa_i + \tau_{B,n}} = B^2 \tag{2.13}$$

where $\mathbf{Q}_i$ is the i^{th} eigenvector of $\mathbf{C}_0$ and here κ_i is the corresponding eigenvalue.

2.4.2. Computation of the projection

Under the alternative, we consider the space generated by the Gaussian measure with covariance function $C_{\mathcal{R}_1}$; recall that $C_{\mathcal{H}^K} = C_{\mathcal{R}_0} + C_{\mathcal{R}_1}$. Denote by $\mathbf{C}_1$ the matrix with (i,j) entry $C_{\mathcal{R}_1}(X_i, X_j)$. We need to project the functions in $\mathcal{R}_1$ onto $\mathcal{R}_0$ and consider the orthogonal part; recall that this ensures that the sample version of the orthogonality condition (2.5) is satisfied. In practice this is amounts to finding the residuals in a ridge regression. We regress each column of $\mathbf{C}_1$ on the columns of $\mathbf{C}_0$. We denote by $\mathbf{C}_1^{(r)}$ the r^{th} column in $\mathbf{C}_1$. Let $\mathbf{S}$ be the diagonal matrix with (i,i) diagonal entry equal to $\partial^2 \ell_{\mu_{0n}}(Z_i)$. We approximately project $\mathbf{C}_1^{(r)}$ onto the column space spanned by $\mathbf{C}_0$ minimizing the loss function

$$\left(\mathbf{C}_1^{(r)} - \mathbf{C}_0 \mathbf{b}^{(r)}\right)^T \mathbf{S} \left(\mathbf{C}_1^{(r)} - \mathbf{C}_0 \mathbf{b}^{(r)}\right) + \rho \left(\mathbf{b}^{(r)}\right)^T \mathbf{C}_0 \mathbf{b}^{(r)}.$$

Here ρ is chosen to go to zero with the sample size (Theorem 1 and Corollary 1). In applications, we may just use a subset of R columns from $\mathbf{C}_1$ and to avoid notational trivialities, say the first R. If there are no computational constraints, we can choose $R = n$. The solution for all $r = 1, 2, ..., R$ is

$$\hat{\mathbf{b}}^{(r)} = \left(\mathbf{C}_0 + \rho \mathbf{S}^{-1}\right)^{-1} \mathbf{C}_1^{(r)}, \tag{2.14}$$

and can be verified substituting it in the first order conditions. Note that (2.14) depends on the the loss function only through $\mathbf{S}$. We define the r^{th} instruments by the $n \times 1$ vector $\hat{\mathbf{h}}^{(r)} = \mathbf{C}_1^{(r)} - \mathbf{C}_0 \hat{\mathbf{b}}^{(r)}$, which is the vector of residuals from the above penalised (Ridge) regression. In sample, when $\rho = 0$, this is orthogonal to the column space of $\mathbf{C}_0$.

2.4.3. The test statistic

Define $\mathbf{e}_0$ to be the $n \times 1$ vector of generalised residuals in the estimation procedure. This means that the i^{th} entry in $\mathbf{e}_0$ is equal to $\partial \ell_{\mu_{0n}}(Z_i)$. The test statistic is $\hat{S}_n = \sum_{r=1}^{R} \left(\hat{\mathbf{e}}_0^T \hat{\mathbf{h}}^{(r)}\right)^2 / R$. Under regularity conditions, if $\mu_0 \in \operatorname{int}(\mathcal{R}_0(B))$, the $R \times 1$ vector

$$\mathbf{g} := \left(\hat{\mathbf{e}}_0^T \hat{\mathbf{h}}^{(1)}, \hat{\mathbf{e}}_0^T \hat{\mathbf{h}}^{(2)}, ..., \hat{\mathbf{e}}_0^T \hat{\mathbf{h}}^{(R)}\right)^T \tag{2.15}$$

is asymptotically Gaussian for any fixed R, and the (k, l) entry of its covariance matrix is consistently estimated by $n^{-1} \sum_{i=1}^{n} \mathbf{e}_{0,i}^2 \hat{\mathbf{h}}_i^{(k)} \hat{\mathbf{h}}_i^{(l)}$. The i^{th} subscript denotes the i^{th} entry in the vector. When the generalised residuals are independent of X, the (k, l) entry in the covariance matrix simplifies to

$$\left(n^{-1} \hat{\mathbf{e}}_0^T \hat{\mathbf{e}}_0\right) \left[n^{-1} \left(\hat{\mathbf{h}}^{(k)}\right)^T \hat{\mathbf{h}}^{(l)}\right].$$

However, this is not assumed here. The distribution of $\hat{S}_n$ can be simulated from the process $\sum_{l=1}^{R} \omega_{n,l} N_l^2$, where the random variables N_l are i.i.d. standard normal and the real valued coefficients $\omega_{n,l}$ are $1/R$ times the eigenvalues of the estimated covariance matrix.

Operational remarks

1. If $C_{\mathcal{R}_1}$ is not explicitly given, we can set $C_{\mathcal{R}_1} = C_{\mathcal{H}^K}$ in the projection step.
2. Instead of $\mathbf{C}_1$ $n \times n$ we can use a subset of the columns of $\mathbf{C}_1$, e.g. $R < n$ columns. Similarly, the r^{th} column of $\mathbf{C}_1$ can be replaced by an $n \times 1$ vector with i^{th} entry $C_{\mathcal{R}_1}(X_i, z_r)$ where z_r is an arbitrary element in $\mathcal{X}^K$. Each column is an instrument. Given that $\mathbf{C}_1$ is singular for n large enough, this does not affect the power of the test because the test functions are highly correlated.

3. To keep the functions $h^{(r)}$ homogeneous before the projection, we should set the r^{th} column of $\mathbf{C}_1$ to have i^{th} entry equal to $C_{\mathcal{H}^K}(X_i, z_r) / \sqrt{C_{\mathcal{H}^K}(z_r, z_r)}$. Note that $h^{(r)}(\cdot) := C_{\mathcal{H}^K}(\cdot, z_r) / \sqrt{C_{\mathcal{H}^K}(z_r, z_r)}$ satisfies $\left|h^{(r)}\right|_{\mathcal{H}^K} = 1$ by the reproducing kernel property.
4. When the series expansion (2.8) for the covariance is known, we can use the elements in the expansion for estimation and testing. For example, suppose $\mathcal{V}_0$ and $\mathcal{V}_1$ are mutually exclusive subsets of the natural numbers such that $C_{\mathcal{R}_j}(s,t) = \sum_{v \in \mathcal{V}_j} \lambda_v \varphi_v(s) \varphi_v(t)$ for $j \in \{0,1\}$. We can directly "project" the elements in $\left\{\lambda_v^{1/2}\varphi_v : v \in \mathcal{V}_1\right\}$ onto the linear span of $\left\{\lambda_v^{1/2}\varphi_v : v \in \mathcal{V}_0\right\}$ by ridge regression with penalty ρ. For $\mathcal{V}_j$ of finite but increasing cardinality, the procedure covers sieve estimators with restricted coefficients. Note that $h^{(r)} = \lambda_r^{1/2}\varphi_r$ satisfies $\left|h^{(r)}\right|_{\mathcal{H}^K} = 1$ for $r \in \mathcal{V}_1$.

Choice of parameters Choice of B in (2.11) can be based on cross-validation, among other methods. In the simulations, for simplicity, we choose B to be a multiple of the sample variance of the data (see the paragraph Estimation Details and Hypotheses, in Sections 4.1, for details). We experimented with the choice of the penalty ρ in the estimation of the sample projection (2.6). In practice, the projection only requires to compute the residuals from a ridge regression, as shown in (2.14). We found that results were not sensitive to ρ. In fact, we found that performing the projection using a pseudo inverse in the computation of (2.14) led to reasonable results, in a number of situations. For this reason, we use this approach in one the simulations. This means computing $\hat{\mathbf{b}}^{(r)} = (\mathbf{C}_0\mathbf{S}\mathbf{C}_0)^{-}\mathbf{C}_0\mathbf{S}\mathbf{C}_1^{(r)}$, where $(\mathbf{C}_0\mathbf{S}\mathbf{C}_0)^{-}$ is the pseudo inverse of $(\mathbf{C}_0\mathbf{S}\mathbf{C}_0)$. For the regression problem under the square error loss, $\mathbf{S}$ can be taken the be the identity matrix. Then, the Ridge regression coefficient for the projection simplifies to $\hat{\mathbf{b}}^{(r)} = (\mathbf{C}_0)^{-}\mathbf{C}_1^{(r)}$.

Additional remarks The procedure can be seen as a J-Test where the instruments are given by the test functions $\hat{\mathbf{h}}^{(r)}$. For a small number of test functions R, a pivotal statistic for our testing procedure can be derived from $\hat{J} := \mathbf{g}'\left[\hat{Var}(\mathbf{g})\right]^{-1}\mathbf{g}$ where $\hat{Var}(\mathbf{g})$ is an estimator of the covaraince matrix of the random vector $\mathbf{g}$ in (2.15). Using the results in Section 3.2 (Theorem 1 and Proposition 2) it is simple to show that the test statistic is asymptotically chi-square with R degrees of freedom. This assumes that $\hat{Var}(\mathbf{g})$ converges to $Var(\mathbf{g})$. In general, the covariance matrix of the vector in (2.15) would be high dimensional (many instruments for large R). Hence it would not be invertible. In finite samples, from a computational point of view, a pseudo inverse could be used in place of $\left[\hat{Var}(\mathbf{g})\right]^{-1}$. However, the convergence of $\hat{J}$ would be to a chi-square with degrees of freedom less than R, a number that would have to be estimated. For this reason, we work directly with the unstandardized statistic. This is common in some high dimensional problems, as it is the case in

functional data analysis.

The test statistic $\hat{S}_n$ has high power when the set of test functions for which the null can be rejected is large. When this not the case, the vector $\mathbf{g}$ in (2.15) may still have some entries that are approximately mean zero. Hence, the average $\hat{S}_n = \sum_{r=1}^{R} \left(\mathbf{e}_0^T \hat{\mathbf{h}}^{(r)}\right)^2 / R$ may obfuscate the departures from the null. In this case, the statistic $\max_{r \leq R} \left|\mathbf{e}_0^T \hat{\mathbf{h}}^{(r)}\right|$ might be preferable. While the maximum of correlated Gaussian random variables can be simulated or approximated, this can be operationally challenging [17, Theorem 3.4]. The rest of the paper provides details and justification for the estimation and testing procedure using the statistics $\hat{S}_n$. The theoretical justification beyond simple heuristics is technically involved. Section 4 (Tables 3 and 5) shows that failing to use the projection procedure discussed in this paper leads to poor inference when we ignore the second term in (2.4).

2.5. Examples of applications

Next we provide a few examples of application of the test statistic to a variety of problems.

Example 3 (Regression)**.** *For the square error loss, we have that* $\ell_\mu(z) = |y - \mu(x)|^2$ *where* $y \in \mathbb{R}$ *and* x *take values in a compact subset of a Euclidean space. Then,* $\partial \ell_\mu(z) = -2(y - \mu(x))$, $\partial^2 \ell_\mu(z) = 2$.

Example 4 (Classification)**.** *For classification, we can use logistic regression:* $\Pr(Y = 1|X = x) = \left(1 + e^{\mu(x)}\right)^{-1}$, $y \in \{0, 1\}$. *The loss function is minus the loglikelihood per observation derived from this probability, i.e.* $\ell_\mu(z) = y\mu(x) + \ln\left(1 + e^{-\mu(x)}\right)$, *after simplifications. Then, defining* $p(x; \mu) := \left(1 + e^{\mu(x)}\right)^{-1}$, $\partial \ell_\mu(z) = (y - p(x; \mu))$, *and* $\partial^2 \ell_\mu(z) = p(x; \mu)(1 - p(x; \mu))$.

Example 5 (Counting)**.** *For arrival counting problems, we can consider the Poisson distribution* $\Pr(Y = y|X = x) = e^{y\mu(x)} \exp\left\{-e^{\mu(x)}\right\} / y!$, $y \in \{0, 1, 2, ...\}$. *We can then define the loss function* $\ell_\mu(z) = -y\mu(x) + e^{\mu(x)}$. *Then,* $\partial \ell_\mu(z) = -\left(y - e^{\mu(x)}\right)$, *and* $\partial^2 \ell_\mu(z) = e^{\mu(x)}$.

Example 6 (Density of failure times)**.** *For the time until an event arrival, consider the survival function* $\Pr(Y > y|X = x) = \exp\left\{-ye^{\mu(x)}\right\}$, $y \in (0, \infty)$. *We define the loss function to be minus the log of the density function:* $ye^{\mu(x)} - \mu(x)$. *We then have* $\partial \ell_\mu(z) = ye^{\mu(x)} - 1$, *and* $\partial^2 \ell_\mu(z) = ye^{\mu(x)}$.

2.6. Examples of restrictions

The following examples focus on restrictions on μ_0 irrespective of the loss function and problem considered. Section 2.5 provided an illustration of such problems.

It is not necessary that $\mathcal{R}_0 \cap \mathcal{R}_1 = \emptyset$, but $\mathcal{R}_0$ must be a proper subspace of $\mathcal{H}^K$ as otherwise there is no restriction to test. In summary, $\mathcal{R}_1$ is not necessarily the complement of $\mathcal{R}_0$ in $\mathcal{H}^K$. A few examples clarify the framework. We shall make use of the results reviewed in Section 2.2 when constructing the covariance functions and in consequence the restrictions.

Example 7. *Consider the additive covariance* $C_{\mathcal{H}^K}(s,t) = \sum_{k=1}^{K} C\left(s^{(k)}, t^{(k)}\right)$ *so that* $\mu(x) = \sum_{k=1}^{K} f^{(k)}\left(x^{(k)}\right)$ *as in (1.2), though* $x^{(k)}$ *could be d-dimensional as in Example 1. Consider the subspace* $\mathcal{R}_0$ *such that* $f^{(1)} = 0$*. This is equivalent to* $C_{\mathcal{R}_0}(s,t) = \sum_{k=2}^{K} C\left(s^{(k)}, t^{(k)}\right)$*. In consequence, we can set* $C_{\mathcal{R}_1}(s,t) = C\left(s^{(1)}, t^{(1)}\right)$.

Some functional restrictions can also be naturally imposed.

Example 8. *Suppose that* $\mathcal{H}^K$ *is an additive space of functions, where each univariate function is an element in the Sobolev Hilbert space of index* V *on* $[0,1]$*, i.e. functions with* V *square integrable weak derivatives. Then,* $C_{\mathcal{H}^K}(s,t) = \sum_{k=1}^{K} C\left(s^{(k)}, t^{(k)}\right)$ *where* $C\left(s^{(k)}, t^{(k)}\right) = \sum_{v=1}^{V-1} \lambda_v^2 \left(s^{(k)} t^{(k)}\right)^v + H_V\left(s^{(k)}, t^{(k)}\right)$ *and where* H_V *is the covariance function of the* V*-fold integrated Brownian motion (see Example 2). Consider the subspace* $\mathcal{R}_0$ *that restricts the univariate RKHS for the first covariate to be the set of linear functions, i.e.* $f^{(1)}\left(x^{(1)}\right) = cx^{(1)}$ *for real c. Then,* $C_{\mathcal{R}_0} = \lambda_1^2 s^{(1)} t^{(1)} + \sum_{k=2}^{K} C\left(s^{(k)}, t^{(k)}\right)$*. Hence we can choose* $C_{\mathcal{R}_1} = \sum_{v=2}^{V-1} \lambda_v^2 \left(s^{(1)} t^{(1)}\right)^v + H_V\left(s^{(1)}, t^{(1)}\right)$.

In the above examples, $\mathcal{R}_1$ is the complement of $\mathcal{R}_0$ in $\mathcal{H}^K$. However, we can just consider spaces $\mathcal{R}_0$ and $\mathcal{R}_1$ to define the model under the null and the space of instruments under the alternative. Hence, no direct reference to $\mathcal{H}^K$ is needed.

Example 9. *Suppose* $C_K(s,t)$ *is a universal kernel on* $[0,1]^K \times [0,1]^K$ *(see Example 1). We suppose that* $C_{\mathcal{R}_0} = \sum_{k=1}^{K} C\left(s^{(k)}, t^{(k)}\right)$*, while* $C_{\mathcal{R}_1} = C_K(s,t)$*. If* C *is continuous and bounded on* $[0,1] \times [0,1]$*, then,* $\mathcal{R}_0 \subset \mathcal{R}_1$*. In this case we are testing an additive model against a general nonlinear one.*

Example 9 fits in the framework of the paper with a slight change of notation. To see this, let $\mathcal{X}^{K+1} = \prod_{k=1}^{K+1} \mathcal{X}^{(k)}$ and $\mathcal{H}^{K+1} = \bigoplus_{k=1}^{K+1} \mathcal{H}^{(k)}$. Here, $\mathcal{H}^{(k)}$ is a RKHS on $\mathcal{X}^{(k)} = [0,1]$ for $k \leq K$, and $\mathcal{H}^{(K+1)}$ is a RKHS on $\mathcal{X}^{(K+1)} = [0,1]^K$. Formally, this also requires us to define $X = \left(X^{(1)}, ..., X^{(K)}, X^{(K+1)}\right)$ with $X^{(K+1)} = \left(X^{(1)}, ..., X^{(K)}\right)$.

The examples above can be extended to test more general models.

Example 10. *Consider the varying coefficients regression function* $\mu(X_i) = bX_i^{(1)} + \beta\left(X_i^{(2)}, ..., X_i^{(K)}\right) X_i^{(1)}$*. The function* $\beta\left(X_i^{(2)}, ..., X_i^{(K)}\right)$ *can be restricted to be linear or additive under the null* $\mu \in \mathcal{R}_0$*. In the additive case,* $C_{\mathcal{R}_0}(s,t) = \lambda_0^2 + s^{(1)} t^{(1)} + \sum_{k=1}^{K} C\left(s^{(k)}, t^{(k)}\right) s^{(1)} t^{(1)}$*. In finance, when* Y *is a continuous response and* $\mathbb{E}(Y|X) = \mu(X)$*, this model can be used to test the conditional Capital Asset Pricing Model. It includes the semiparametric model discussed in [10].*

3. Asymptotic analysis

3.1. Regularity conditions

Throughout the paper, $\lesssim$ means that the l.h.s. is bounded by an absolute constant times the r.h.s..

Condition 1. *The set $\mathcal{H}$ is a RKHS on a compact subset of a separable Banach space $\mathcal{X}$, with continuous uniformly bounded kernel C admitting an expansion (2.8). Furthermore, in (2.8), $\lambda_v^2 \lesssim v^{-2\eta}$ with exponent $\eta > 1$ and with linearly independent continuous uniformly bounded functions $\varphi_v : \mathcal{X} \to \mathbb{R}$. If each additive component has a different covariance kernel, the condition is meant to apply to each of them individually.*

Condition 2. *The sequence $(Z_i)_{i\in\mathbb{Z}}$ ($Z_i = (Y_i, X_i)$) is independent identically distributed (i.i.d.).*

Recall the loss $L(z,t)$ from Section 2.1. Define $c_K := \max_{s\in\mathcal{X}^K} \sqrt{C_{\mathcal{H}^K}(s,s)}$ and let $\bar{B} := c_K B$. Define $\Delta_k(z) := \max_{|t|\leq \bar{B}} \left|\partial^k L(z,t)/\partial t^k\right|$ for $k = 0, 1, 2, \ldots$. Let P_x be the law of Z given $X = x$. We shall consider two competing regularity conditions on the loss. The first is a conditional form of Bartlett identity which says that $P_x \partial \ell_{\mu_0}^2 = P_x \partial^2 \ell_{\mu_0}$. The other requires a uniform lower bound on the second derivative of the intensity: $\inf_{z,t} d^2 L(z,t)/dt^2 > 0$ for $z \in \mathcal{Z}$ and $t \in \left[-\bar{B}, \bar{B}\right]$. We refer to this as strict convexity condition.

Condition 3. *The loss $L(z,t)$ is non-negative, twice continuously differentiable for real t in an open set containing $\left[-\bar{B}, \bar{B}\right]$, and satisfies either the above Bartlett identity or the strict convexity condition. Moreover, $P\Delta_0 + P\Delta_1^{2p} + P\Delta_1^p \Delta_2^p + \left|P_x \partial^2 \ell_{\mu_0}\right|_\infty + \left|P_x \Delta_3\right|_\infty < \infty$ for some $p > 2$.*

Condition 4. *The estimator in (2.11) is such that $|\mu_{0n} - \mu_0|_2 = O_P(n^{-\alpha})$ for some $\alpha > 1/4$.*

We shall refer to the above as the Regularity Conditions.

3.1.1. Remarks on conditions

Condition 1 A minimal decay condition for the coefficients λ_v would be $\lambda_v \lesssim v^{-\eta}$ with $\eta > 1/2$ as this is essentially required for $\sum_{v=1}^\infty \lambda_v^2 \varphi_v^2(s) < \infty$ for any $s \in \mathcal{X}$. Here, we require that the condition is strengthened to $\eta > 1$, at the very least. To put this into perspective, note that the covariance in Example 1 satisfies Condition 1 with exponentially decaying coefficients λ_v [21, Chapter 4.3.1]; the covariance in Examples 2 and 8 satisfies $\lambda_v \lesssim v^{-\eta}$ with η equal to the number of derivatives of μ.

Condition 1 restricts the individual covariances in $C_{\mathcal{H}^K}$. The same condition is inherited by the individual covariances that comprise $C_{\mathcal{R}_0}$ (i.e. Condition 1 applies to each individual component of $C_{\mathcal{R}_0}$). In a similar vein, in Example 9, the covariance $C_{\mathcal{R}_1}$ can be seen as the individual covariance of a multivariate

variable $X^{(K+1)} := \left(X^{(1)}, ..., X^{(K)}\right)$ and $C_{\mathcal{R}_1}$ will have to satisfy (2.8) where the functions φ_v are on $\mathcal{X}^{(K+1)}$.

Condition 2 We only consider i.i.d. data. We could allow for dependent data as done by other authors (e.g. [23]). This would require additional technical conditions and lead to distracting technicalities.

Condition 3 All the loss functions in Section 2.5 and many more satisfy Condition 3, using the fact that $|\mu|_\infty \leq \bar{B}$. Recall that $\bar{B}$ was defined just before Condition 3. Note that the loss in Example 6 does not satisfy the uniform lower bound on the second derivative of the loss, but is does satisfy the Bartlett identity when $\mu_0 \in \mathcal{H}^K(B)$. On the other hand, loss functions that are not smooth, such as the ones used to derive conditional quantiles would not. Such loss functions require to impose smoothness directly on $P\partial\ell_\mu h$. This extension would come at the cost of additional conditions and requires a separate treatment beyond the scope of the paper.

Condition 4 The estimator in (2.11) is a nuisance parameter. Its rate of convergence can be very slow, depending on the problem. The following special case of Theorem 3 in Sancetta (2021) gives rates of convergence. The proof in Sancetta (2021) uses the strict convexity condition in Condition 3 for identification ([23, remarks just after Equation (19)]). This is unnecessary and can be replaced by the milder conditional convexity condition $\inf_x \mathbb{E}[\inf_t d^2 L(Z,t)/dt^2 | X = x] > 0$ for $x \in \mathcal{X}^K$ and $t \in \left[-\bar{B}, \bar{B}\right]$. Written in another way, this is $\inf_x P_x \inf_\mu \partial^2 \ell_\mu > 0$, $\mu \in \mathcal{R}_0(B)$ and $x \in \mathcal{X}^K$. This is satisfied by the loss in Example 6.

Proposition 1. *Suppose that the Regularity Conditions, the conditional convexity condition and* $\mu_0 \in \text{int}(\mathcal{R}_0(B))$ *hold. Then,* $|\mu_{0n} - \mu_0|_2 = O_P(\min\left\{n^{-\alpha}, n^{-1/4}\right\})$, *where* $\alpha = \frac{\gamma}{2}\left(\frac{2\eta-1}{2\eta+(\gamma-1)}\right)$ *and* $\gamma := \left(\frac{p-1}{p}\right)$.

The convergence rate in Proposition 1 can become arbitrarily close to the parametric one when $\eta \to \infty$. For example, for univariate functions that have an arbitrarily large number of η square integrable derivatives (see Example 2). The rate in Proposition 1 is not optimal, but it is sufficient for most practical purposes. For example, [36, Theorem 1.9] for optimal rates in the case of the univariate regression problem of smooth functions estimation under the square error loss.

The nuisance parameter does not need to be an exact minimizer as in (2.11). An asymptotic minimizer suffices ([31, Theorem 3.2.5]). In some cases the exact minimizer in an RKHS can be difficult to compute in practice, and approximations are needed ([21], [3, Chapter 8], [19], [23], and references therein). For simplicity, we do not pursue this here.

3.2. Test statistic

Recall that $\mathcal{R}_0(B) := \mathcal{R}_0 \cap \mathcal{H}^K(B)$ for any $B > 0$ and similarly for $\mathcal{R}_1(B)$. Suppose that μ_0 in (2.1) lies in the interior of $\mathcal{R}_0(B)$. Then, the moment equation $P\partial\ell_{\mu_0} h = 0$ holds for any $h \in \mathcal{R}_1$. This is because, by definition of (2.1), $\partial\ell_{\mu_0}$ is orthogonal to all elements in $\mathcal{H}^K$. By linearity, one can restrict attention to $h \in \mathcal{R}_1(1)$ (i.e. $\mathcal{R}_1(B)$ with $B = 1$). For such functions h, the statistic $P_n\partial\ell_{\mu_0} h$ is normally distributed [23, Theorem 4]. In practice, μ_0 is rarely known and it is replaced by μ_{0n} in (2.11). The estimator μ_{n0} does not need to satisfy $P_n\partial\ell_{\mu_{0n}} h = 0$ for any h in $\mathcal{H}^K(1)$ under the null. Moreover, the nuisance parameter affects the asymptotic distribution.

For fixed $\rho \geq 0$, let Π_ρ be the penalized population projection operator such that

$$\Pi_\rho h = \arg \inf_{\nu \in \mathcal{R}_0} P\partial\ell_{\mu_0}^2 (h - \nu)^2 + \rho |\nu|_{\mathcal{H}^K}^2 \tag{3.1}$$

for any $h \in \mathcal{H}^K$. Let the population projection operator be Π_0, i.e. (3.1) with $\rho = 0$. Given that Π_0 is unknown, we replace it with the sample projection operator in (2.6). To ease notation, write $\Pi_n = \Pi_{n,\rho}$ for $\rho = \rho_n$. Let the symbol $\asymp$ mean that the l.h.s. is bounded above and below by absolute constants times the r.h.s.. The following holds.

Theorem 1. *Suppose that the Regularity Conditions hold and that $\mu_0 \in \text{int}(\mathcal{R}_0(B))$. Moreover, suppose that in (2.6) $\rho \asymp n^{-\theta}$ with $\theta < \alpha$ and $2\alpha + \theta > 1$, where α is as in Condition 4. Then,*

$$P_n\partial\ell_{\mu_{0n}}(h - \Pi_n h) \to G(h - \Pi_0 h), \; h \in \mathcal{H}^K(1),$$

weakly, where the r.h.s. is a mean zero Gaussian process with covariance function

$$\Sigma(h, h') := \mathbb{E}G(h - \Pi_0 h) G(h' - \Pi_0 h') = P\partial\ell_{\mu_0}^2 (h - \Pi_0 h)(h' - \Pi_0 h') \tag{3.2}$$

for any $h, h' \in \mathcal{H}^K(1)$.

An estimator for (3.2) is required in order to fully approximate the limiting process in Theorem 1. A consistent estimator for Σ at $h, h' \in \mathcal{H}^K(1)$ is given by Σ_n such that

$$\Sigma_n(h, h') = P_n\partial^2\ell_{\mu_{0n}}(h - \Pi_n h)(h' - \Pi_n h'). \tag{3.3}$$

Proposition 2. *Suppose that the Regularity Conditions hold and that $\rho \asymp n^{-\theta}$ with $\theta < \alpha$. Then,*

$$\sup_{h,h' \in \mathcal{H}^K(1)} |\Sigma_n(h, h') - \Sigma(h, h')| \to 0$$

in probability.

Theorem 1 says that the score statistic that uses $(h - \Pi_n h)$ in place of h converges to a Gaussian process. Hence, we can construct test statistics using any

continuous map of the empirical process $\left\{P_n \partial \ell_{\mu_{0n}}(h - \Pi_n h) : h \in h \in \mathcal{H}^K(1)\right\}$. A computationally feasible statistic is obtained defining a finite set $\tilde{\mathcal{R}}_1 \subseteq \mathcal{R}_1 \cap \mathcal{H}^K(1)$. Let the cardinality of $\tilde{\mathcal{R}}_1$ be R, for definiteness. For the sake of clarity in what follows, fix an order in the elements in $\tilde{\mathcal{R}}_1$. Recall that the test statistic in (2.7) and Section 2.4.3 is

$$\hat{S}_n := \frac{1}{R} \sum_{h \in \tilde{\mathcal{R}}_1} \left[P_n \partial \ell_{\mu_{0n}}(h - \Pi_n h)\right]^2. \tag{3.4}$$

By Theorem 1 and the continuous mapping theorem, this converges in distribution to the random variable

$$S := \frac{1}{R} \sum_{h \in \tilde{\mathcal{R}}_1} \left[G(h - \Pi_0 h)\right]^2. \tag{3.5}$$

The distribution of S is standard, but depends on Σ. Let ω_k be the k^{th} scaled eigenvalue of the covariance matrix $\left\{\Sigma(h, h') : h, h' \in \tilde{\mathcal{R}}_1\right\}$, i.e., $\omega_k \psi_k(h) = \frac{1}{R} \sum_{h' \in \tilde{\mathcal{R}}_1} \Sigma(h, h') \psi_k(h')$, where the k^{th} eigenvector $\left\{\psi_k(h) : h \in \tilde{\mathcal{R}}_1\right\}$ satisfies $\frac{1}{R} \sum_{h \in \tilde{\mathcal{R}}_1} \psi_k(h) \psi_l(h) = 1$ if $k = l$ and zero otherwise.

Proposition 3. *The random variable in (3.5) can be written as* $S = \sum_{k \geq 1} \omega_k N_k^2$, *where the random variables* N_k *are independent standard normal. The equality holds in* L_2.

Remark 1. *Given that* R *is finite, we can just compute the eigenvalues (in the usual sense) of the matrix with entries* $\Sigma(h, h')/R$, $h'h' \in \tilde{\mathcal{R}}_1$.

Let $\hat{S} := \sum_{k \geq 1} \omega_{nk} N_k^2$ where ω_{nk} is the k^{th} scaled eigenvalue of the covariance matrix $\left\{\Sigma_n(h, h') : h, h' \in \tilde{\mathcal{R}}_1\right\}$ (see Remark 1). The random variable $\hat{S}$ does not depend on any unknown quantities. Hence, we can simulate from it. Next, we show that asymptotically $\hat{S}_n$, $\hat{S}$ and S have same distribution. In practice this means that in large sample, $\Pr\left(\hat{S}_n > x\right) \to \Pr(S > x)$ and $\Pr\left(\hat{S} > x\right) \to \Pr(S > x)$ under the null hypothesis that $\mu_0 \in \operatorname{int}(\mathcal{R}_0(B))$. P-values for the Type I error of the test statistic $\hat{S}_n$ can be consistently obtained simulating from $\hat{S}$ and computing the empirical probability of observing values greater than $\hat{S}_n$.

Theorem 2. *Under the conditions of Theorem 1,* $\hat{S}_n$ *and* $\hat{S}$ *converge in distribution to* S.

Recall that we denote by P_x the law of Z given $X = x$. Define the function $w : \mathcal{X}^K \to \mathbb{R}$ such that $w(x) := P_x \partial^2 \ell_{\mu_0}$. The function w might be known under the null. In this case, $\partial^2 \ell_{\mu_{0n}}$ in (2.6) can be replaced by w, i.e., define the empirical projection as the arg inf of

$$P_n w (h - \nu)^2 + \rho |\nu|^2_{\mathcal{H}^K} = \frac{1}{n} \sum_{i=1}^{n} w(X_i) (h(X_i) - \nu(X_i))^2 + \rho |\nu|^2_{\mathcal{H}^K} \tag{3.6}$$

w.r.t. $\nu \in \mathcal{R}_0$. For example, for the regression problem, using the square error loss, w is constant and can be set to one with no loss of generality.

Corollary 1. *Suppose the function w is known. Replace $\Pi_n h$ with the minimizer of (3.6) in the construction of the test statistic $\hat{S}_n$ and Σ_n.*

Suppose that the conditions of Theorem 1 hold with $\theta < 1/2$ instead of $\theta < \alpha$. Then, Theorems 1 and 2 continue to hold. Similarly, suppose that the conditions of Theorem 2 hold with $\theta < 1/2$ instead of $\theta < \alpha$. Then, Theorem 2 continues to hold.

Corollary 1 improves on Theorem 2 as it imposes less restrictions on the penalty ρ. Despite the technicalities required to justify the procedure, the implementation shown in Section 2.4 is straightforward. In fact, $\partial \ell_{\mu_{0n}}$ evaluated at $Z_i := (Y_i, X_i)$ is the score for the i^{th} observation and it is the i^{th} entry in $\hat{\mathbf{e}}_0$, as defined in Section 2.4. On the other hand the vector $\hat{\mathbf{h}}^{(r)}$ has i^{th} entry $\left(h^{(r)}(X_i) - \Pi_n h^{(r)}(X_i)\right)$ and $\tilde{\mathcal{R}}_1 = \left\{h^{(1)}, ..., h^{(R)}\right\}$. For example we can set

$$\tilde{\mathcal{R}}_1 = \left\{C_{\mathcal{R}_1}(\cdot, z_r) : z_r \in \mathcal{X}^K, r = 1, 2, ..., R\right\}.$$

We conclude noting that Theorem 1 holds uniformly in $h \in \mathcal{H}^K(1)$. Hence, $\tilde{\mathcal{R}}_1$ is only restricted by considerations pertaining to power and computational constraints. In particular it can be chosen to have more weight on the direction of certain alternatives if the researcher has prior information.

4. Finite sample evidence via simulation examples

We use some simulation examples to shed further light on the importance of the projection procedure. First, we consider an example where the nuisance parameter is potentially high dimensional. In the second example, the nuisance parameter is infinite dimensional. In both cases, not using the projection to account for the nuisance parameter leads to the wrong inference. In particular, we compute the statistics in (2.3) and (2.7) and use the asymptotic distribution of $\sqrt{n} P_n \partial \ell_{\mu_0} h$ to find the p-values. In what follows, we refer to (2.3) as the statistic that ignores the nuisance parameter.

4.1. Finite dimensional model

We consider a high dimensional problem in the sense that the number of estimated parameters can be of the same order of magnitude as the sample size. We also consider simplifications exploiting the known structure of the covariance series expansion (2.8). This is to show that the framework is applicable to common problems such as polynomial regression with constrained coefficients.

Simulation design: True models Consider the regression problem where $Y_i = \mu_0(X_i) + \varepsilon_i$, the number of covariates $X^{(k)}$ is $K = 10$, and the sample size is $n = 100$, and 1000. The covariates are i.i.d. standard Gaussian random variables that are then truncated to the interval $\mathcal{X} = [-2, 2]$. Before truncation, the cross-sectional correlation between $X^{(k)}$ and $X^{(l)}$ is $\varrho^{|k-l|}$ with $\varrho \in \{0, 0.75\}$, $k, l = 1, 2, ..., K$. The error terms are i.i.d. mean zero, Gaussian with variance

TABLE 1

True Models. List of true models and short name for ease of reference are defined.

Name		
Lin3	$\mu_0(X) = \sum_{k=1}^{3} b_k X^{(k)}$	$b_k = 1/3$, $k = 1, 2, 3$.
LinAll	$\mu_0 := \sum_{k=1}^{10} b_k X_i^{(k)}$	$b_k = 1/10$, $k = 1, 2, ..., 10$.
Lin1Poly4	$\mu_0(X) = X^{(1)} + \sum_{v=1}^{9} b_{4,v} \left(X^{(4)}/2\right)^v$	$b_{4,v}$ uniformly distributed in $[-20/v, 20/v]$, $v = 1, 2, ..., 9$.

TABLE 2

Models in Restricted and Test Spaces. List of models in $\mathcal{R}_0$ and $\mathcal{R}_1$ and their short name for ease of reference are defined.

Names	
Lin1	$C_{\mathcal{R}_0}(s,t) = s^{(1)}t^{(1)}$
Lin2	$C_{\mathcal{R}_0}(s,t) = \sum_{k=1}^{2} s^{(k)}t^{(k)}$
Lin3	$C_{\mathcal{R}_0}(s,t) = \sum_{k=1}^{3} s^{(k)}t^{(k)}$
LinAll	$C_{\mathcal{R}_0}(s,t) = \sum_{k=1}^{10} s^{(k)}t^{(k)}$
Lin1Poly	$C_{\mathcal{R}_0}(s,t) = s^{(1)}t^{(1)} + \sum_{k=2}^{10} C\left(s^{(k)}, t^{(k)}\right)$
	$C_{\mathcal{R}_1}(s,t) = C_{\mathcal{H}^{10}} - C_{\mathcal{R}_0}(s,t)$ for all of the above
	$C_{\mathcal{H}^{10}}(s,t) = \sum_{k=1}^{10} C\left(s^{(k)}, t^{(k)}\right)$, where $C\left(s^{(k)}, t^{(k)}\right) = \sum_{v=1}^{10} v^{-2.2}\left(s^{(k)}t^{(k)}\right)^v$

such that the signal to noise ratio $\sigma^2_{\mu/\varepsilon}$ is equal to 1 and 0.2. This is equivalent to an R^2 of 0.5 and 0.167, i.e. a moderate and low R^2.

The specifications for μ_0 together with a short name are defined in Table 1 for ease of reference.

In Lin1Poly4 the first variable enters the model linearly, the fourth variable enters it in a nonlinear fashion, while the remaining variables do not enter the model. The choice of random coefficient for Lin1Poly4 is to mitigate the dependence on a specific nonlinear functional form. The number of simulations is 1000.

Note that for Lin1Poly4, we choose the fourth covariate to enter the model nonlinearly in order to make the problem harder when covariates are correlated with a Toeplitz correlation. In this case, the test functions are more correlated with the nuisance parameter than if we had chosen the second covariate to enter the true model nonlinearly.

Estimation details and hypotheses We let $\mathcal{H}^{10}$ be generated by a polynomial additive kernel. The details are in Table 2 for ease of reference. The covariance kernel $C_{\mathcal{H}^{10}}$ is such that the true models in Table 1 all lie in a strict subset of $\mathcal{H}^{10}$.

Recall the definition of $\mathcal{L}^K(B)$ at the end of Section 2.2. Estimation is carried out in $\mathcal{L}^{10}(B)$ using a greedy algorithm with number of iterations equal to 500 [23, Section 4.1]. This approach speeds up calculations. It also allow us to assess whether there is a distortion in the test results when the estimator minimizes the objective function only approximately.

The parameter B is chosen equal to $10\hat{\sigma}_Y$ where $\hat{\sigma}_Y$ is the sample standard deviation of Y, which is a crude approach to keep simulations manageable.

The eigenvalues from the sample covariance were used to simulate the limiting process from which the p-values were derived using 10^4 simulations.

The restricted models are again shown in Table 2. The restrictions are chosen to be able to show coverage probabilities when the restriction is true, as well as the power of the test when the restriction is false. In all cases we test against the full unrestricted model with kernel $C_{\mathcal{H}^{10}}(s,t)$.

Test functions We can exploit the structure of the covariance kernel and simplify the testing procedure. We can define the functions $h^{(v,k)}: \mathcal{X}^K \to \mathbb{R}$ such that $h^{(v,k)}(s) = v^{-1.1}\left(s^{(k)}\right)^v$. Then, $\left|h^{(v,k)}\right|_{\mathcal{H}^K} = 1$. Moreover, the functions in $\mathcal{R}_1$ are in the linear span of such test functions. For example, when $\mathcal{R}_0$ is Lin1Poly (see Table 2), the functions in $\mathcal{R}_0$ are in the linear span of $\left\{h^{(1,1)}\right\} \cup \left\{h^{(v,k)}: v \leq 10, k = 2,3,...,10\right\}$.

To estimate the projection, we use the pseudo-inverse rather than a penalty $\rho > 0$ (see Section 2.4). This allow us to see if this simple and crude approach is viable.

Results Table 3 reports the frequency of rejections for a given nominal size of the test. Here, results are for $n = 1000$, a signal to noise level $\sigma^2_{\mu/\varepsilon} = 1$, and correlation $\varrho = 0$ under the three different true designs: Lin3, LinAll, and Lin1Poly4. The column heading "No Π" means that no correction was used in estimating the test statistic (i.e. test statistic ignoring the presence of nuisance parameters). The results for the other configurations of sample size, signal to noise ratio and correlation in the variables were similar. The Lin1Poly model is only estimated when the true model is Lin1Poly4. Here, we only report a subset of the tested hypotheses (Lin3 and LinAll, only). The complete set of results is in Section B in the Appendix. Without using the projection adjustment, the size of the test can be highly distorted, as expected. The results reported in Table 3 show that the test (properly constructed using the projection adjustment) has coverage probability relatively close to the nominal one when the null holds, and that the test has a good level of power. Here, we report the case of uncorrelated covariates. For correlated covariates the distortion is much more prominent even in low dimensions (see the full set of results in Section B, in the Appendix).

4.2. Infinite dimensional model

Simulation design: True model Consider a bivariate regression model with independent standard normal errors. The regression function is

$$\text{Lin1Poly2: } \mu_0(x) = b\left(\frac{1}{2}x^{(1)} + \frac{3}{2}x^{(2)} - 4\left(x^{(2)}\right)^2 + 3\left(x^{(2)}\right)^3\right), \tag{4.1}$$

where the scalar coefficient b is chosen so that the signal to noise ratio is 1 and 0.2 and $x \in \mathcal{X}^2$ where $\mathcal{X} = [-2,2]$. The covariates X_i and the errors ε_i together with the other details are as in Section 4.1. We shall refer to (4.1) as Lin1Poly2.

TABLE 3

Finite Dimensional Model. Simulated frequency of rejections for $n = 1000$, signal to noise ratio $\sigma^2_{\mu/\varepsilon} = 1$, and variables correlation $\varrho = 0$. Results for different true models are reported. For true model Lin3, no restriction should be rejected. For true model LinAll, restriction Lin3 should be rejected. For true model Lin1Poly4, restrictions Lin3 and LinAll should be rejected. The column heading "Size" stands for the nominal size.

Size	Lin3		LinAll		Lin1Poly	
	No Π	Π	No Π	Π	No Π	Π
True model: Lin3						
0.10	0.09	0.11	0.07	0.10	-	-
0.05	0.05	0.05	0.04	0.06	-	-
True model: LinAll						
0.10	1.00	1.00	0.49	0.08	-	-
0.05	1.00	1.00	0.23	0.05	-	-
True model: Lin1Poly4						
0.10	1.00	1.00	0.92	0.91	0.03	0.1
0.05	1.00	1.00	0.9	0.88	0.02	0.05

TABLE 4

Models in Restricted and Test Spaces. List of models in $\mathcal{R}_0$ and $\mathcal{R}_1$ and their short name for ease of reference are defined

Names	
Lin1NonLin2	$C_{\mathcal{R}_0}(s,t) = 0.5\left(1 + \sum_{k=1}^{2} s^{(k)}t^{(k)}\right) + 0.5\exp\left\{\frac{1}{2}\left(\frac{s^{(2)}-t^{(2)}}{0.75}\right)^2\right\}$
	$C_{\mathcal{R}_1}(s,t) = 0.5\exp\left\{-\frac{1}{2}\left(\frac{s^{(1)}-t^{(1)}}{0.75}\right)^2\right\}$
LinAll	$C_{\mathcal{R}_0}(s,t) = \sum_{k=1}^{10} s^{(k)}t^{(k)}$
	$C_{\mathcal{R}_1}(s,t) = 0.5\exp\left\{-\frac{1}{2}\left[\sum_{k=1}^{2}\left(\frac{s^{(k)}-t^{(k)}}{0.75}\right)^2\right]\right\}$

TABLE 5

Infinite Dimensional Model. Simulated frequency of rejections for $n = 1000$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The true model is Lin1Poly2. Restriction LinAll should be rejected. The column heading "Size" stands for the nominal size.

			Lin1NonLin2		LinAll	
ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π
0	1	0.10	0.00	0.09	0.99	1.00
0	1	0.05	0.00	0.04	0.83	1.00
0	0.2	0.10	0.00	0.09	0.00	1.00
0	0.2	0.05	0.00	0.04	0.00	1.00
0.75	1	0.10	0.00	0.09	1.00	1.00
0.75	1	0.05	0.00	0.03	1.00	1.00
0.75	0.2	0.10	0.00	0.09	0.13	1.00
0.75	0.2	0.05	0.00	0.03	0.01	1.00

Estimation details and hypotheses We consider two hypotheses as shown in Table 4. The hypothesis Lin1NonLin2 postulate a linear model for the first covariate and a nonlinear for the second. The true model μ_0 is in $\mathcal{R}_0$, hence this hypothesis allows us to verify the size of a Type I error. In the case of the hypothesis LinAll, the true model is not in $\mathcal{R}_0$ and this hypothesis allows us to

verify the power of the test.

The nuisance parameter is computed using exact estimation in $\mathcal{H}^2(B)$ solving the problem as if it were a Ridge regression problem with Ridge penalty $\tau_{B,n}$ such that the constraint is satisfied, as shown in (2.13). All the other details are as in Section 4.1.

Test functions We let the test functions $h^{(r)}: \mathcal{X}^2 \to \mathbb{R}$ be such that $h^{(r)}(s) = C_{\mathcal{R}_1}(s, X_r)/\sqrt{C_{\mathcal{R}_1}(X_r, X_r)}$, $r = 1, 2, ..., n$. We project on the functions $\{C_{\mathcal{R}_0}(\cdot, X_r) : i = 1, 2, ..., n\}$ using Ridge regression as shown in (2.14). The penalty ρ in the projection is chosen equal to $\rho_n := \left[\frac{1}{n}\sum_{r=1}^{n} h^{(r)}(X_r)\right]/n^{1/2}$. To speed up the test, we then randomly sample 100 of such test functions when $n > 100$. Note that ρ_n is still computed using the original set of n functions. Finally, given that the class of functions we are estimating are infinite differentiable, we expect η to be greater than any finite integer (see Proposition 1 and the discussion following it). In consequence, we can choose $\rho_n \asymp n^{-\theta}$ with θ arbitrarily close to 1/2 (Corollary 1).

Results Table 5 reports the frequency of rejections for $n = 1000$. The complete set of results is in Section B in the Appendix. The results show a considerable improvement relative to the naive test. The additional cost of implementing the projections is marginal, as we just need to compute residuals from a Ridge regression.

Given that the test functions are exponential functions, the test statistic for "No Π" is equivalent to "Model specification test 1" in [5]. There, only an upper bound for the Type I error probability was derived, under a parametric null hypothesis [5, Theorem 6]. Projecting away the impact of the nuisance parameter, we are able to obtain the asymptotic distribution of the test statistic even when the model under the null is nonparametric (Lin1NonLin2).

4.3. Local power example

We further investigate the power of the test. To this end, it is informative to gauge an idea of its local power. While a theoretical analysis can be difficult, we can resort on simulations. We focus on a modification of the infinite dimensional setup of the previous section.

Simulation design: True model Consider a bivariate regression model with independent standard normal errors. The regression function is

$$\text{Lin1Poly2Local: } \mu_0(x) = b\mu^{(1)}\left(x^{(1)}\right) + c_n\mu^{(2)}\left(x^{(2)}\right), \tag{4.2}$$

$$\mu^{(1)}\left(x^{(1)}\right) = \frac{1}{2}x^{(1)},\ \mu^{(2)}\left(x^{(2)}\right) = \frac{3}{2}x^{(2)} - 4\left(x^{(2)}\right)^2 + 3\left(x^{(2)}\right)^3$$

$c_n = \left(\hat{\sigma}^2 n\right)^{-1/2} c$, $c \in \{0.1, 0.5, 1, 2, 5, 10\}$, where $\hat{\sigma}$ is set equal to the sample standard deviation of $\mu^{(2)}\left(X^{(2)}\right)$, the scalar coefficient b is chosen so that the

TABLE 6

Local Infinite Dimensional Model. Simulated frequency of rejections for $n = 1000$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$ for $\mu^{(1)}$, and variables correlation ϱ. The true model is Lin1Poly2Local. Restriction LinAll is tested and should be rejected. The column heading "Size" stands for the nominal size. The local power is reported for different deviations from the lower dimensional model via the constant c defined after (4.2).

ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π	No Π	Π
					c			
			0.1		0.5		1	
0	1	0.10	0.00	0.09	0.00	0.11	0.00	0.18
0	1	0.05	0.00	0.04	0.00	0.06	0.00	0.11
0	0.2	0.10	0.00	0.09	0.00	0.09	0.00	0.10
0	0.2	0.05	0.00	0.04	0.00	0.05	0.00	0.06
0.75	1	0.10	0.00	0.09	0.00	0.11	0.00	0.21
0.75	1	0.05	0.00	0.05	0.00	0.06	0.00	0.12
0.75	0.2	0.10	0.00	0.08	0.00	0.09	0.00	0.10
0.75	0.2	0.05	0.00	0.04	0.00	0.05	0.00	0.06
					c			
			2		5		10	
0	1	0.10	0.00	0.40	0.00	0.98	0.26	1.00
0	1	0.05	0.00	0.27	0.00	0.95	0.02	1.00
0	0.2	0.10	0.00	0.16	0.00	0.47	0.00	0.95
0	0.2	0.05	0.00	0.09	0.00	0.33	0.00	0.89
0.75	1	0.10	0.00	0.53	0.01	1.00	0.94	1.00
0.75	1	0.05	0.00	0.39	0.00	0.99	0.62	1.00
0.75	0.2	0.10	0.00	0.19	0.00	0.61	0.00	0.99
0.75	0.2	0.05	0.00	0.10	0.00	0.48	0.00	0.98

signal to noise ratio of $\mu^{(1)}$ is 1 and 0.2 and $x \in \mathcal{X}^2$ where $\mathcal{X} = [-2, 2]$. By this we mean that $P\left|b\mu^{(1)}\right|^2$ is equal to 1 and 0.2 times the variance of the error ε_i, which is set to one. The covariates X_i, the errors ε_i and other details are as in Section 4.2. We shall refer to (4.2) as Lin1Poly2Local. As we increase c, (4.2) deviates from $b\mu^{(1)}\left(x^{(1)}\right)$. However, the deviation is local because $c_n = O\left(n^{-1/2}\right)$.

Estimation details and hypotheses The estimation details are the same as in Section 4.2. To gauge the local power of the test for different values of c, we test the restricted model LinAll in Table 4.

Test functions The details are as in Section 4.2 for the test functions corresponding to LinAll in Table 4.

The results in Table 6 show that the power increases with c. However, failing to account for the nuisance parameter greatly reduces the power of the test.

5. Extension to an additional nuisance parameter

The approach of this paper can be extended to include additional nuisance parameters. For simplicity we discuss the case of one additional real valued nuisance parameter that we denote by β. Consider a loss $L(z, t, s) : \mathcal{Z} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

Then, $\ell_{\mu,\beta}(z) = L(z, \mu(x), \beta(x))$. The parameters μ and β could depend on possibly different subsets of $x \in \mathcal{X}^K$. The goal is to test restrictions on μ while we have no interest in β. If a plug-in estimator β_n for the true parameter β_0 can be obtained, we would minimize the profile loss $P_n \ell_{\mu,\beta_n}$ w.r.t. μ and obtain μ_{0n}. Then, we would carry out the test procedure of this paper with no change.

Estimation of β_0 would not affect the distribution of the test statistic, under standard conditions. We outline such conditions. Let

$$\partial^{(k,l)} \ell_{\mu,\beta}(z) = \partial^{k+l} L(z,t,s) / \left(\partial t^k \partial s^l\right)$$

setting $t = \mu(x)$ and $s = \beta(x)$ after differentiation. Then, we require the stochastic equicontinuity condition

$$\sqrt{n}(P_n - P)\left(\partial^{(1,0)} \ell_{\mu_{0n},\beta_n} - \partial^{(1,0)} \ell_{\mu_0,\beta_0}\right) h = o_p(1) \tag{5.1}$$

and the orthogonality condition $\sqrt{n} P \partial^{(1,0)} \ell_{\mu_{0n},\beta_n} h = o_p(1)$. This is the same setup as in andrews94 with the complications induced by testing restrictions on μ_0. Recall that, by definition, $P\partial^{(1,0)} \ell_{\mu_0,\beta_0} h = 0$ as μ_0 solves the first order conditions. Then, $P\partial^{(1,0)} \ell_{\mu_{0n},\beta_n} h$ is equal to

$$\begin{aligned} P\partial^{(1,0)} \ell_{\mu_{0n},\beta_n} h - P\partial^{(1,0)} \ell_{\mu_0,\beta_0} h = & P\partial^{(2,0)} \ell_{\mu_0,\beta_0}(\mu_{0n} - \mu_0) h \\ & + P\partial^{(1,1)} \ell_{\mu_0,\beta_0}(\beta_n - \beta_0) h + o_p(1). \end{aligned}$$

The above display holds if $\left|P_x \partial^{(1,2)} \ell_{\mu,\beta}\right|_\infty < \infty$ uniformly in μ and β and $|\beta_n - \beta_0|_2 = o_p\left(n^{-1/4}\right)$. This is deduced by standard arguments ([2, Equations 4.12-4.13]; see also [23, Equation 25]). The methodology of the paper is not affected by estimation of β_0 if $P\partial^{(1,1)} \ell_{\mu_0,\beta_0}(\beta_n - \beta_0) h = 0$, in the above display. This orthogonality condition holds rather frequently. It should be contrasted with $P\partial^{(2,0)} \ell_{\mu_0,\beta_0}(\mu_{0n} - \mu_0) h = 0$, which, as we know does not hold in general.We give an illustrative example in the next section.

5.1. Example: Classification with inverse probability weighting

We consider the nonlinear binary classification problem using logistic regression (see Section 2.5). However, we allow for missing data. Suppose that R_i is an indicator with value equal to one if the data is observed and zero otherwise. We suppose that $\Pr(R_i = 1|X_i) = \beta_0(X_i)$. We assume a random sample from $Z = (Y, X)$ where $Y = \left(Y^{(1)}, Y^{(2)}\right) \in \{0,1\}^2$ with binary response $Y^{(1)}$, and $Y_i^{(2)} = R_i$. To account for missing data, we can use inverse probability weighting of the data. Hence, we use the loss

$$P_n \ell_{\mu,\beta} = \frac{1}{n} \sum_{i=1}^n \left(Y_i^{(2)} / \beta(X_i)\right) \left[Y_i^{(1)} \mu(X_i) + \ln\left(1 + e^{-\mu(X_i)}\right)\right].$$

The functions μ and β may depend on different subsets of X. The quantity in the square brackets can be replaced by other losses for other type of problems. The framework is the one of [35]. Assume that under the null the generalised error $\varepsilon_i := Y_i^{(1)} - \left(1 + e^{\mu_0(X_i)}\right)^{-1}$ is mean zero conditioning on X_i and $Y_i^{(2)}$. Then, using the fact that $\partial^{(1,0)}\ell_{\mu_0,\beta_0}(Z_i) = \left(Y_i^{(2)}/\beta(X_i)\right)\varepsilon_i$ (see Section 2.5), we have that

$$P\partial^{(1,1)}\ell_{\mu_0,\beta_0}(\beta - \beta_0)h = \mathbb{E}\frac{Y_i^{(2)}\varepsilon_i}{\beta^2(X_i)}(\beta(X_i) - \beta_0(X_i))h(X_i) = 0$$

for any β bounded away from zero and $h \in \mathcal{H}^K$. For this problem the nuisance parameter β satisfies the orthogonality condition in [2], as expected because the expectation of ε_i conditional on X_i is zero. Hence, the testing procedure of this paper is not unaffected.

We conclude noting that additional conditions need to be imposed in order to show the distributional results of the paper. These details can be derived mechanically accounting for the additional parameter β_n in the proofs.

6. Conclusion

This paper considers the problem of testing subspace restrictions for possibly additive models in reproducing kernel Hilbert spaces. As well known, the presence of a high dimensional nuisance parameter invalidates standard asymptotic inference. We show how to remove the dependence on the nuisance parameter and recover standard asymptotic arguments. This is achieved constructing test functions that are essentially orthogonal to the nuisance parameter. In practice this only requires to run a ridge regression and compute its residuals. Hence, the test is simple to implement. Simulation results show that failing to carry out this approach in the presence of nuisance parameters can invalidate inference even for relatively simple problems.

Appendix A: Proofs

Recall that $\ell_\mu(Z) = L(Z, \mu(X))$ and $\partial^k \ell_\mu(Z) = \partial^k L(Z,t)/\partial t^k\big|_{t=\mu(X)}$, $k \geq 1$. Condition 3 implies Fréchet differentiability of $P\ell_\mu$ and $P\partial\ell_\mu$ (w.r.t. $\mu \in \mathcal{H}^K$) at μ in the direction of $h \in \mathcal{H}^K$. It can be shown that these two derivatives are $P\partial\ell_\mu h$ and $P\partial^2\ell_\mu hh$, respectively. For this purpose, we view $P\ell_\mu$ as a map from the space of uniformly bounded functions on $\mathcal{X}^K$ to $\mathbb{R}$. The details can be derived following the steps in [26, proof of Lemma 2.21] or [15, proof of Lemma A.4]. The application of those proofs to the current scenario, essentially requires that the loss function $L(Z,t)$ is differentiable w.r.t. real t, and that μ is uniformly bounded, together with integrability of the quantities Δ_0, and Δ_1, as implied by Condition 3. It will also be necessary to take the Fréchet derivative of $P_n\ell_\mu$ and $P_n\partial\ell_\mu h$ conditioning on the sample data. By Condition 3

this will also hold because Δ_0, and Δ_1 are finite. Following the aforementioned remarks, when the loss function is three times differentiable, we also have that for any $h \in \mathcal{H}^K$, the Fréchet derivative of $P\partial^2 \ell_\mu h$ in the direction of $h' \in \mathcal{H}^K$ is $P\partial^3 \ell_\mu hh'$. These facts will be used throughout the proofs with no further mention. Moreover, throughout, for notational simplicity, we tacitly suppose that $\sup_{x\in\mathcal{X}^K} \sqrt{C_{\mathcal{H}^K}(x,x)} = 1$ so that $h \in \mathcal{H}^K(B)$ implies that $|h|_\infty \leq B$ for any $B > 0$. This follows from the reproducing kernel property (e.g. [23, Lemma 1]).

A.1. Entropy numbers

Denote by $N\left(\epsilon, \mathcal{F}, |\cdot|_p\right)$ the ϵ-covering number of a set $\mathcal{F}$, relative to the L_p norm. This is the minimum number of open balls of L_p radius ϵ needed to cover $\mathcal{F}$. The entropy is the logarithm of the covering number. Denote by $N_{[]}\left(\epsilon, \mathcal{F}, |\cdot|_p\right)$ the ϵ-bracketing number, relative to the L_p norm, of the set $\mathcal{F}$. This is the minimum number of L_p ϵ-brackets needed to cover $\mathcal{F}$. Given two functions $f_L \leq f_U$ such that $|f_L - f_U|_p \leq \epsilon$, an L_p ϵ-bracket $[f_L, f_U]$ is the set of all functions $f \in \mathcal{F}$ such that $f_L \leq f \leq f_U$. The covering and bracketing number relative to the uniform norm coincide. We have the following ϵ-entropy estimates.

Lemma 1. *Under Condition 1,*

1. $\ln N_{[]}\left(\epsilon, \mathcal{H}^K(B), |\cdot|_\infty\right) \lesssim K(B/\epsilon)^{2/(2\eta-1)}$;

2. $\ln N_{[]}\left(\epsilon, \mathcal{F}, |\cdot|_p\right) \lesssim K(B/\epsilon)^{2/(2\eta-1)}$ *for* $\mathcal{F} := \left\{\partial \ell_\mu h : \mu \in \mathcal{H}^K(B), h \in \mathcal{H}^K(1)\right\}$ *and any* $p \in [1,\infty]$ *satisfying Condition 3;*

3. $\ln N_{[]}\left(\epsilon, \mathcal{F}, |\cdot|_p\right) \lesssim K(B/\epsilon)^{2/(2\eta-1)}$ *for* $\mathcal{F} := \{\partial \ell_\mu^2 hh' : \mu \in \mathcal{H}^K(B), h, h' \in \mathcal{H}^K(1)\}$ *and any* $p \in [1,\infty]$ *satisfying Condition 3.*

Proof. Points 1. and 2. are, respectively, Lemma 3 and 4 in [23]. Hence, we only prove Point 3, which is proved similarly. By Condition 3 and the triangle inequality, for $h, h', g, g' \in \mathcal{H}^K(1)$, we have that

$$\left|\partial \ell_\mu^2 hh' - \partial \ell_{\mu'}^2 gg'\right| \leq \left|\partial \ell_\mu^2 - \partial \ell_{\mu'}^2\right| \sup_{h\in\mathcal{H}^K(1)} |h|^2 + \sup_{\mu\in\mathcal{H}^K(B)} \left|\partial \ell_\mu^2\right| (|hh' - gh'| + |gh' - gg'|).$$

By Condition 3, we have the following bounds $\left|\partial \ell_\mu^2(z)\right| \leq \Delta_1^2(z)$ and $\left|\partial \ell_\mu^2(z) - \partial \ell_{\mu'}^2(z)\right| \leq 2\Delta_1(z)\Delta_2(z)|\mu(x) - \mu'(x)|$. Moreover, $|h|_\infty \leq 1$ for $h \in \mathcal{H}^K(1)$. By these remarks, the previous display is bounded by

$$2\Delta_1\Delta_2(z)|\mu - \mu'|_\infty + \Delta_1^2(z)\left(|h - g|_\infty + |h' - g'|_\infty\right).$$

Theorem 2.7.11 in [31] says that the L_p ϵ-bracketing number of class of functions satisfying the above Lipschitz kind of condition is bounded by the L_∞ ϵ'-covering number of $\mathcal{H}^K(B) \times \mathcal{H}^K(1)$ with $\epsilon' = 4\epsilon\left(P\left|\Delta_1^{2p} + \Delta_1^p \Delta_2^p\right|\right)^{-1/p}$. □

A.2. Preliminary lemmas

Lemma 2. *Under the Regularity Conditions,* $\sup_{h\in\mathcal{H}^K(\delta)} P_n\partial\ell_{\mu_{0n}}h = O_P(n^{-1/2}B\delta)$ *for* μ_{n0} *as in (2.11).*

Proof. The first order condition for the penalised sample estimator $\mu_{n,\tau}$ reads

$$P_n\partial\ell_{\mu_{n,\tau}}h = -2\tau\left\langle\mu_{n,\tau},h\right\rangle_{\mathcal{H}^K} \leq 2\tau\left|\mu_{n,\tau}\right|_{\mathcal{H}^K}|h|_{\mathcal{H}^K} \tag{A.1}$$

for any $h\in\mathcal{H}^K(\delta)$, $\delta<\infty$. In consequence, $\sup_{h\in\mathcal{H}^K(\delta)} P_n\partial\ell_{\mu_{n,\tau}}h \leq 2\tau\left|\mu_{n,\tau}\right|_{\mathcal{H}^K}\delta$. From, [23, Theorem 2], there is a $\tau_{B,n} = O_P\left(n^{-1/2}\right)$ such that $\mu_{n,\tau} = \mu_{0n}$ when $\tau = \tau_{B,n}$. Hence, $\sup_{h\in\mathcal{H}^K(\delta)} P_n\partial\ell_{\mu_{0n}}h = O\left(n^{-1/2}B\delta\right)$. □

Lemma 3. *Under the Regularity Conditions,* $\left|\mu_{0n}-\mu_0\right|_\infty \to 0$ *in probability.*

Proof. The Regularity Conditions assume convergence in L_2. To turn the L_2 convergence into uniform, note that $\mathcal{H}^K(B)$ is compact under the uniform norm by Lemma 1 (Point 1.). Moreover, functions in $\mathcal{H}^K(B)$ are continuous and defined on a compact domain $\mathcal{X}^K$. In consequence, any convergent sequence in $\mathcal{H}^K(B)$ converges uniformly. □

For any positive finite measure Q on $\mathcal{Z}$, write $\nu_{Q,\rho} = \arg\inf_{\nu\in\mathcal{R}_0} Q\left(h-\nu\right)^2 + \rho\left|\nu\right|_{\mathcal{H}^K}$.

Lemma 4. *Suppose that* Q *and* P *are positive finite measures on* $\mathcal{Z}$*. Then,*

$$\left|\nu_{Q,\rho}-\nu_{P,\rho}\right|_{\mathcal{H}^K} \leq \frac{2}{\rho}\sum_{v=1}^{\infty}\lambda_v\left|\left(Q-P\right)\left(h-\nu_{P,\rho}\right)\varphi_v\right|$$

Moreover, if $\nu_{Q,\rho}$ *is bounded for any* $\rho\to 0$*, then* $\left|\nu_{Q,\rho}-\nu_Q\right|_{\mathcal{H}^K}\to 0$ *and similarly for* $\nu_{P,\rho}$*.*

Proof. By [26, Corollary 5.10] applied to the square error loss,

$$\left|\nu_{Q,\rho}-\nu_{P,\rho}\right|_{\mathcal{H}^K} \leq \frac{2}{\rho}\left|Q\left(h-\nu_{P,\rho}\right)\Phi - P\left(h-\nu_{P,\rho}\right)\Phi\right|_{\mathcal{H}^K}, \tag{A.2}$$

where $\Phi(x) = C_{\mathcal{H}^K}(\cdot,x)$ is the canonical feature map. By (2.8), the canonical feature map can be written as $\Phi(x) = \sum_{v=1}^{\infty}\lambda_v^2\varphi_v(\cdot)\varphi_v(x)$. This implies that,

$$\left(Q-P\right)\left(h-\nu_{P,\rho}\right)\Phi(x) = \sum_{v=1}^{\infty}\left[\lambda_v^2\left(Q-P\right)\left(h-\nu_{P,\rho}\right)\varphi_v\right]\varphi_v(x).$$

By (2.10), and the above,

$$\begin{aligned}\left|\left(Q-P\right)\left(h-\nu_{P,\rho}\right)\Phi\right|_{\mathcal{H}^K}^2 &= \sum_{v=1}^{\infty}\frac{\left[\lambda_v^2\left(Q-P\right)\left(h-\nu_{P,\rho}\right)\varphi_v\right]^2}{\lambda_v^2}\\ &= \sum_{v=1}^{\infty}\lambda_v^2\left[\left(Q-P\right)\left(h-\nu_{P,\rho}\right)\varphi_v\right]^2.\end{aligned}$$

Bounding the r.h.s. of (A.2) by the square root of the above display, and using the triangle inequality for each term in the sum, we deduce the first statement in the lemma. The last statement in the lemma is continuity w.r.t. ρ and corresponds to [26, Theorem 5.17]. □

A.3. Convergence of projection operators

We use the operators Π_ρ, $\Pi_{n,\rho}$, $\tilde{\Pi}_{n,\rho}$ such that for any $h \in \mathcal{H}^K$:

$$\begin{aligned}
\Pi_\rho h &:= \arg \inf_{\nu \in \mathcal{R}_0} P\partial^2 \ell_{\mu_0} (h-\nu)^2 + \rho |\nu|^2_{\mathcal{H}^K} \text{ as in (3.1)} \\
\Pi_{n,\rho} h &:= \arg \inf_{\nu \in \mathcal{R}_0} P_n^2 \partial \ell_{\mu_{0n}} (h-\nu)^2 + \rho |\nu|^2_{\mathcal{H}^K} \text{ as in (2.6)} \\
\tilde{\Pi}_{n,\rho} h &:= \arg \inf_{\nu \in \mathcal{R}_0} P\partial \ell^2_{\mu_{0n}} (h-\nu)^2 + \rho |\nu|^2_{\mathcal{H}^K} . \qquad \text{(A.3)}
\end{aligned}$$

To ease notation, we may write $\Pi_n = \Pi_{n,\rho}$ when $\rho = \rho_n$. Moreover, recall that we also write Π_0 for Π_ρ with $\rho = 0$.

The proof uses some preliminary results. In what follows, we may assume that $K = 1$. This is to avoid notational complexities that could obscure the main steps in the derivations. Because of additivity, this is not restrictive as long as K is bounded.

Lemma 5. *Suppose that $h \in \mathcal{H}^K(1)$. Then, $|\Pi_0 h|_{\mathcal{H}^K} \leq 1$.*

Proof. By construction, the linear projection $\Pi_0 h$ satisfies $\Pi_0 h \in \mathcal{R}_0$ and $\Pi_0 (h - \Pi_0 h) = 0$. Hence, the space $\mathcal{H}^K$ is the direct sum of the set $\mathcal{R}_0$ and its complement in $\mathcal{H}^K$, say $\mathcal{R}_0^c$. These sets are orthogonal. Note that we do not necessarily have $\mathcal{R}_0^c = \mathcal{R}_1$ unless the basis that spans $\mathcal{R}_1$ is already linearly independent of $\mathcal{R}_0$. By Lemma 9.1 in [32], $|h|_{\mathcal{H}^K} = |\Pi_0 h|_{\mathcal{R}_0} + |h - \Pi_0 h|_{\mathcal{R}_0^c}$. The norms are the ones induced by the inner products in the respective spaces. But, $|\Pi_0 h|_{\mathcal{R}_0^c} = 0$. Hence, we have that $|\Pi_0 h|_{\mathcal{R}_0} = |\Pi_0 h|_{\mathcal{H}^K} \leq |h|_{\mathcal{H}^K} = 1$. □

Lemma 6. *Under the Regularity Conditions, if $\rho n^\alpha \to \infty$, then,*

$$\sup_{h \in \mathcal{H}^K(1)} \left| \left(\Pi_\rho - \tilde{\Pi}_{n,\rho} \right) h \right|_{\mathcal{H}^K} \to 0$$

in probability.

Proof. Let $\tilde{P}$ and $\tilde{P}_n$ be finite positive measures on $\mathcal{Z}$ such that $d\tilde{P}/dP = \partial^2 \ell_{\mu_0}$ and $d\tilde{P}_n/dP = \partial^2 \ell_{\mu_{0n}}$. By Lemma 4,

$$\left| \left(\Pi_\rho - \tilde{\Pi}_{n,\rho} \right) h \right|_{\mathcal{H}^K} \leq \frac{2}{\rho} \sum_{v=1}^{\infty} \lambda_v \left| \left(\tilde{P}_n - \tilde{P} \right) (h - \Pi_\rho h) \varphi_v \right| . \qquad \text{(A.4)}$$

Taking derivatives, we bound each term in the absolute value by

$$\left| P \left(\partial^2 \ell_{\mu_{0n}} - \partial^2 \ell_{\mu_0} \right) (h - \Pi_\rho h) \varphi_v \right| \leq \left| P \sup_{\mu \in \mathcal{H}^K(B)} \left| \partial^3 \ell_\mu \right| (\mu_{0n} - \mu_0) (h - \Pi_\rho h) \varphi_v \right| .$$

By Lemma 5 and the definition of penalized estimation, $|\Pi_\rho h|_{\mathcal{H}^K} \leq |\Pi_0 h|_{\mathcal{H}^K} \leq 1$ uniformly in $\rho \geq 0$. Hence, $|h - \Pi_\rho h|_\infty \leq 2$. Moreover, $\sup_{v \geq 1} |\varphi_v|_\infty < \infty$. Therefore, the r.h.s. of the above display is bounded, independently of ρ, by a constant multiple of

$$\sqrt{P \sup_{\mu \in \mathcal{H}^K(B)} |\partial^3 \ell_\mu|^2} \sqrt{P |\mu_{0n} - \mu_0|^2} = \sqrt{P\Delta_3^2}\, |\mu_{0n} - \mu_0|_2$$

The term $P\Delta_3^2$ is finite by Condition 3. By assumption, we have that $|\mu_{0n} - \mu_0|_2 = O_p(n^{-\alpha})$. Using the above display to bound (A.4), deduce that the lemma holds true if $\rho^{-1} n^{-\alpha} = o_p(1)$, as stated in the lemma. Taking supremum w.r.t. $h \in \mathcal{H}^K(1)$ in the above steps, deduce that the result holds uniformly in $h \in \mathcal{H}^K(1)$. □

Lemma 7. *Under the Regularity Conditions, we have that* $\sup_{h \in \mathcal{H}^K(1)} \left| \left(\Pi_{n,\rho} - \tilde{\Pi}_{n,\rho} \right) h \right|_{\mathcal{H}^K} \to 0$ *in probability for any* ρ *such that* $\rho n^{1/2} \to \infty$ *in probability.*

Proof. Following the same steps as in the proof of Lemma 6, deduce that

$$\left| \left(\Pi_{n,\rho} - \tilde{\Pi}_{n,\rho} \right) h \right|_{\mathcal{H}^K} \leq \frac{2}{\rho} \sum_{v=1}^{\infty} \lambda_v \left| (P_n - P)\, \partial^2 \ell_{\mu_{0n}} \left(h - \tilde{\Pi}_{n,\rho} h \right) \varphi_v \right|. \tag{A.5}$$

Each term in the absolute value on the r.h.s. is bounded in L_1 by

$$\begin{aligned} &\mathbb{E} \sup_{h \in \mathcal{H}^K(1), \mu \in \mathcal{H}^K(B), \nu \in \mathcal{H}^K(1)} \left| (P_n - P)\, \partial^2 \ell_\mu (h - \nu) \varphi_v \right| \\ \leq &2\mathbb{E} \sup_{h \in \mathcal{H}^K(1), \mu \in \mathcal{H}^K(B)} \left| (P_n - P)\, \partial^2 \ell_\mu h \varphi_v \right|. \end{aligned}$$

Define the class of functions $\mathcal{F} := \left\{ \partial^2 \ell_\mu h \varphi_k : \mu \in \mathcal{H}^K(B), h \in \mathcal{H}^K(1) \right\}$. Given that φ_v is uniformly bounded, by Lemma 1 it can be deduced that $\mathcal{F}$ is Donsker ([31, Theorem 2.5.6]). Hence the expectation in the above display is $O\left(n^{-1/2}\right)$. By permanence of the Donsker property for any fixed convex combination of Donsker classes ([31, Theorem 2.10.1], the expectation of (A.5) is $O\left(n^{-1/2}\rho^{-1}\right)$. □

In what follows, recall that we defined P_x to be the law of Z conditioning on $X = x$.

Lemma 8. *Suppose that the Regularity Conditions hold. Then, for* ρ *such that* $\rho n^\alpha \to \infty$ *in probability, and for* $n \to 0$*, we have that*

$$\sup_{h \in \mathcal{H}^K(1)} |(\Pi_\rho - \Pi_{n,\rho}) h|_{\mathcal{H}^K} = o_p(1),$$

and

$$\sup_{h \in \mathcal{H}^K(1)} \left| \sqrt{n} P_n \partial \ell_{\mu_{n0}} (\Pi_\rho - \Pi_{n,\rho}) h \right| = o_p(1). \tag{A.6}$$

Finally, if $w(x) = P_x \partial^2 \ell_{\mu_0}$ *is a known function, and* $\Pi_{n,\rho}$ *constructed as in (3.6), the above displays hold for* ρ *such that* $\rho n^{1/2} \to \infty$ *in probability.*

Proof. By the triangle inequality

$$\sup_{h \in \mathcal{H}^K(1)} |(\Pi_\rho - \Pi_{n,\rho}) h|_{\mathcal{H}^K} \leq \sup_{h \in \mathcal{H}^K(1)} \left| \left(\Pi_\rho - \tilde{\Pi}_{n,\rho}\right) h \right|_{\mathcal{H}^K} + \sup_{h \in \mathcal{H}^K(1)} \left| \left(\tilde{\Pi}_{n,\rho} - \Pi_{n,\rho}\right) h \right|_{\mathcal{H}^K}. \tag{A.7}$$

The first statement in the lemma follows by showing that the r.h.s. of the above is $o_p(1)$. This is the case by application of Lemmas 6 and 7.

By the established convergence in $|\cdot|_{\mathcal{H}^K}$, for any $h \in \mathcal{H}^K(1)$, $|(\Pi_\rho - \Pi_{n,\rho}) h|_{\mathcal{H}^K} \leq \delta$ with probability going to one for any $\delta > 0$. Therefore, to prove (A.6), we can restrict attention to a bound for

$$\lim_{\delta \to 0} \sup_{|h|_{\mathcal{H}^K} \leq \delta} \sqrt{n} P_n \partial \ell_{\mu_{0n}} h.$$

By Lemma 2, we know that the above display is zero.

Finally, to show the last statement in the lemma, note that it is Lemma 6 that puts an additional constraint on ρ. However, saying that the function w is known, effectively amounts to saying that we can replace μ_{0n} with μ_0 in the definition of $\tilde{\Pi}_{n,\rho}$ in (A.3). This means that $\tilde{\Pi}_{n,\rho} = \Pi_\rho$ so that the second term in (A.7) is exactly zero and we do not need to use Lemma 6. Therefore, ρ is only constrained according to Lemma 7. □

We also need to bound the distance between Π_ρ and Π_0, but this cannot be achieved in probability under the operator norm.

Lemma 9. *Under the Regularity Conditions, we have that*

$$\sup_{h \in \mathcal{H}^K(1)} P \partial^2 \ell_{\mu_0} (\Pi_\rho h - \Pi_0 h)^2 \leq \rho.$$

Proof. As in the proof of Lemma 6, let $\tilde{P}$ such that $d\tilde{P}/dP = \partial^2 \ell_{\mu_0}$. At first we show that

$$\tilde{P} (\Pi_\rho h - \Pi_0 h)^2 \leq \tilde{P} (h - \Pi_\rho h)^2 - \tilde{P} (h - \Pi_0 h)^2. \tag{A.8}$$

To see this, expand the r.h.s. of (A.8), add and subtract $2\tilde{P}(\Pi_0 h)^2$, and verify that the r.h.s. of (A.8) is equal to

$$-2\tilde{P} h \Pi_\rho h + 2\tilde{P} (h - \Pi_0 h) \Pi_0 h + \tilde{P} \left[(\Pi_\rho h)^2 + (\Pi_0 h)^2 \right].$$

However, $(h - \Pi_0 h)$ is orthogonal, w.r.t. $\tilde{P}$, to elements in $\mathcal{R}_0$, in the sense of (A.3) with $\rho = 0$. Since $\Pi_0 h \in \mathcal{R}_0$, the middle term in the above display is zero. Then, add and subtract $2\tilde{P} \Pi_\rho h \Pi_0 h$ and rearrange to deduce that the above display is equal to

$$2\tilde{P} (\Pi_0 h - h) \Pi_\rho h + \tilde{P} (\Pi_\rho h - \Pi_0 h)^2.$$

Given that $(\Pi_0 h - h)$ is orthogonal to elements in $\mathcal{R}_0$ and $\Pi_\rho h \in \mathcal{R}_0$, the first term on the above display is also zero. This shows that (A.8) holds true. The r.h.s. of (A.8) is bounded as follows,

$$\begin{aligned}\tilde{P}\left(h-\Pi_\rho h\right)^2-\tilde{P}\left(h-\Pi_0 h\right)^2 &\leq \left[\tilde{P}\left(h-\Pi_\rho h\right)^2+\rho\left|\Pi_\rho h\right|_{\mathcal{H}^K}^2\right]-\tilde{P}\left(h-\Pi_0 h\right)^2 \\ &\leq \left[\tilde{P}\left(h-\Pi_0 h\right)^2+\rho\left|\Pi_0 h\right|_{\mathcal{H}^K}^2\right]-\tilde{P}\left(h-\Pi_0 h\right)^2 \\ &= \rho\left|\Pi_0 h\right|_{\mathcal{H}^K}^2\end{aligned}$$

because $|\Pi_\rho h|_{\mathcal{H}^K}$ is positive and $\Pi_\rho h$ is the minimizer of the penalized population loss function (see (A.3)). Using Lemma 5, the r.h.s. of the above display is bounded by ρ. Hence the r.h.s. of (A.8) is bounded by ρ uniformly in $h \in \mathcal{H}^K(1)$, and the lemma is proved. □

The penalised population projection operator is continuous.

Lemma 10. *Under the Regularity Conditions, we have that*

$$P\partial^2\ell_{\mu_0}\sqrt{n}\left(\mu_{0n}-\mu_0\right)\left(\Pi_0 h-\Pi_\rho h\right)=o_p(1) \tag{A.9}$$

for any ρ such that $n^{(1-2\alpha)}\rho \to 0$ in probability.

Proof. By Holder inequality, the absolute value of the l.h.s. of (A.9) is bounded above by

$$\sqrt{n}\left[P\partial^2\ell_{\mu_0}\left(\mu_{0n}-\mu_0\right)^2\right]^{1/2}\left[P\partial^2\ell_{\mu_0}\left(\Pi_0 h-\Pi_\rho h\right)^2\right]^{1/2}. \tag{A.10}$$

Recall that P_x is the law of Z conditional on $X = x$. Hence we can write $P = PP_x$. By the Regularity Conditions, $\left|P_x\partial^2\ell_{\mu_0}\right|_\infty < \infty$, so that

$$\sqrt{n}\left[P\partial^2\ell_{\mu_0}\left(\mu_{0n}-\mu_0\right)^2\right]^{1/2} \lesssim \sqrt{n}\left[P\left(\mu_{0n}-\mu_0\right)^2\right]^{1/2} = O_p\left(n^{(1-2\alpha)/2}\right)$$

using the assumption on convergence of μ_{0n}. Hence, by Lemma 9, deduce that (A.10) is bounded above by a quantity $O_p\left(n^{(1-2\alpha)/2}\rho^{1/2}\right) = o_p(1)$ for the given choice of ρ. □

We can relate the expectation of the squared score to the expectation of the second derivative of the loss. This will needed for an application of the projection correction.

Lemma 11. *Under Condition 3, $P\partial\ell_{\mu_0}^2\left|h'-h\right| \lesssim \sqrt{P\partial^2\ell_{\mu_0}\left(h'-h\right)^2}$.*

Proof. At first, suppose that the Bartlett identity holds. Using $P = PP_x$ and the the aforementioned identity, $P\partial\ell_{\mu_0}^2\left|h'-h\right| = P\partial^2\ell_{\mu_0}\left|h'-h\right|$. Then, by Holder inequality applied to $P\sqrt{\partial^2\ell_{\mu_0}}\sqrt{\partial^2\ell_{\mu_0}\left|h'-h\right|^2}$ we deduce the result because $P\partial^2\ell_{\mu_0}$ is bounded.

Now suppose that the uniform lower bound $\inf_{z,t} d^2 L(z,t)/dt^2 > 0$ holds. Then, multiplying and dividing by $\sqrt{\partial^2 \ell_{\mu_0}}$ and using Holder inequality,

$$P\partial \ell_{\mu_0}^2 |h' - h| \leq \sqrt{P\left(\partial \ell_{\mu_0}^4 / \partial^2 \ell_{\mu_0}\right)} \sqrt{P\partial^2 \ell_{\mu_0} (h' - h)^2}.$$

The uniform lower bound and the moment condition on $\partial \ell_{\mu_0}^4$ ensure that the quantity in the first square root is finite. □

Finally, we show convergence in distribution when there is no nuisance parameter. This is needed, as we shall show that the projection correction is asymptotically equivalent to this.

Lemma 12. *Suppose that the Regularity Conditions hold and that* $\mu_0 \in \text{int}\left(\mathcal{H}^K(B)\right)$. *If* $\rho \to 0$,

$$\sqrt{n} P_n \partial \ell_{\mu_0} (h - \Pi_\rho h) \to G(h - \Pi_0 h), \; h \in \mathcal{H}^K(1),$$

weakly, where the r.h.s. is a mean zero Gaussian process with covariance function

$$\Sigma(h, h') := \mathbb{E} G(h - \Pi_0 h) G(h' - \Pi_0 h') = P\partial \ell_{\mu_0}^2 (h - \Pi_0 h)(h' - \Pi_0 h')$$

for any $h, h' \in \mathcal{H}^K(1)$.

Proof. Any mean zero Gaussian process $(G(h))$ – not necessarily the one in the lemma – is continuous w.r.t. the pseudo norm $d(h, h') = \sqrt{\mathbb{E}|G(h) - G(h')|^2}$ [1, Lemma 1.3.1]. Hence, $d(h, h') \to 0$ implies that $G(h) - G(h') \to 0$ in probability. By Lemma 5, deduce that $(h - \Pi_\rho h) \in \mathcal{H}^K(2)$. Hence, consider the mean zero Gaussian process $(G(h))$ with covariance function $\mathbb{E} G(h) G(h') = P\partial \ell_{\mu_0}^2 hh'$ with $h \in \mathcal{H}^K(2)$. By direct calculation, and this remark on h and h',

$$d^2(h, h') = P\partial \ell_{\mu_0}^2 h (h - h') + P\partial \ell_{\mu_0}^2 h' (h' - h) \lesssim P\partial \ell_{\mu_0}^2 |h' - h|. \quad \text{(A.11)}$$

By Lemma 11 the r.h.s. is bounded above by a constant multiple of $\sqrt{P\partial^2 \ell_{\mu_0} (h - h')^2}$. Hence, to check continuity of the Gaussian process at arbitrary $h \to h'$, we only need to consider $P\partial^2 \ell_{\mu_0} (h - h')^2 \to 0$. We shall use this remark momentarily.

Now, note that by Theorem 4 in [23], which also holds for any $h \in \mathcal{H}^K(2)$, $\sqrt{n} P_n \partial \ell_{\mu_0} h$ converges weakly to a Gaussian process $G(h)$, $h \in \mathcal{H}^K(2)$. Hence, $\sqrt{n} P_n \partial \ell_{\mu_0} (h - \Pi_\rho h)$ converges weakly to $G(h - \Pi_0 h)$ if for any $h \in \mathcal{H}^K(1)$

$$\sup_{h \in \mathcal{H}^K(1)} \lim_{\rho \to 0} |G(h - \Pi_\rho h) - G(h - \Pi_0 h)| = 0$$

in probability. By the initial remarks about continuity, it is sufficient to check that $\sup_{h \in \mathcal{H}^K(1)} P\partial^2 \ell_{\mu_0} (\Pi_0 h - \Pi_\rho h)^2 \to 0$ in probability as $\rho \to 0$. This is the case by Lemma 9. □

A.4. Convergence of sample eigenvalues

We need to estimate the eigenvalues ω_k in order to compute critical values. The following will also prove Proposition 2.

Lemma 13. *Under the conditions of Lemma 8 the following hold in probability:*

1. $\sup_{h,h' \in \mathcal{H}^K(1)} |\Sigma_n(h,h') - \Sigma(h,h')| \to 0$, *where* Σ *and* Σ_n *are as in (3.2) and (3.3), respectively;*

2. $\sup_{k>0} |\omega_{nk} - \omega_k| \to 0$, *where* ω_{nk} *and* ω_k *are the* k^{th} *eigenvalues of the covariance functions with entries* $\Sigma_n(h,h')$ *and* $\Sigma(h,h')$, $h,h' \in \tilde{\mathcal{R}}_1$ *for* $\tilde{\mathcal{R}}_1$ *a countable subset of* $\mathcal{R}_1$, *not necessarily finite;*

3. The population eigenvalues ω_k *are summable and for any* $c \geq 1 + \sum_{k=1}^{\infty} \omega_k$, *we have that* $\Pr\left(\sum_{k=1}^{\infty} \omega_{n,k} > c\right) = o(1)$.

Proof. To show Point 1, use the triangle inequality to deduce that

$$\begin{aligned} |\Sigma_n(h,h') - \Sigma(h,h')| \quad \leq \quad & \left|(P_n - P)\left(\partial \ell^2_{\mu_{0n}}\right)(h - \Pi_n h)(h' - \Pi_n h')\right| \\ & + \left|P\left(\partial \ell^2_{\mu_{0n}} - \partial \ell^2_{\mu_0}\right)(h - \Pi_n h)(h' - \Pi_n h')\right| \\ & + \left|P\partial \ell^2_{\mu_0}(\Pi_0 h - \Pi_n h)(h' - \Pi_n h')\right| \\ & + \left|P\partial \ell^2_{\mu_0}(h - \Pi_0 h)(\Pi_0 h' - \Pi_n h')\right|. \end{aligned} \tag{A.12}$$

It is sufficient to bound each term individually uniformly in $h,h' \in \mathcal{H}^K(1)$.

To bound the first term in (A.12), note that, with probability going to one, $|h - \Pi_n h|_{\mathcal{H}^K} \leq 2 + \epsilon$ for any $\epsilon > 0$ uniformly in $h \in \mathcal{H}^K(1)$, by Lemmas 5 and 8, as $n \to \infty$. By this remark, to bound the first term in probability, it is enough to bound $\left|(P_n - P)\partial \ell^2_{\mu} h h'\right|$uniformly in $\mu \in \mathcal{H}^K(B)$ and $h,h' \in \mathcal{H}^K(2+\epsilon)$. By Lemma 1 (Point 3.) the class of functions is Donsker ([31, Theorem 2.5.6]. We can deduce that this term is $O_p\left(n^{-1/2}\right)$.

To bound the second term in (A.12), note that $P\partial \ell^2_{\mu}$ is Fréchet differentiable w.r.t. μ. To see this, one can use the same arguments as in [26, proof of Lemma 2.21] as long as $P \sup_{\mu \in \mathcal{H}^K(B)} \left|\partial \ell_{\mu} \partial^2 \ell_{\mu}\right| < \infty$, which is the case by the assumptions in the lemma. Hence,

$$\begin{aligned} & \left|P\left(\partial \ell^2_{\mu_{0n}} - \partial \ell^2_{\mu_0}\right)(h - \Pi_n h)(h' - \Pi_n h')\right| \\ \leq \quad & 2\left|P\partial \ell_{\mu_0} \partial^2 \ell_{\mu_0}(\mu_{0n} - \mu_0)(h - \Pi_n h)(h' - \Pi_n h')\right| + o_p(1) \end{aligned}$$

using the fact that $|\mu_{0n} - \mu_0|_{\infty} = o_p(1)$ by Lemma 3. By an application of Lemma 8, again, a bound in probability for the above is given by a bound for

$$2 \sup_{h,h' \in \mathcal{H}^K(2+\epsilon)} \left|P\partial \ell_{\mu_0} \partial^2 \ell_{\mu_0}(\mu_{0n} - \mu_0) h h'\right|.$$

By Lemma 3 and $P\left|\partial \ell_{\mu_0} \partial^2 \ell_{\mu_0}\right| \leq P\Delta_1 \Delta_2 < \infty$, we deduce that the above is $o_p(1)$.

The third term in (A.12) is bounded by

$$P\left|\partial \ell^2_{\mu_0}(\Pi_0 h - \Pi_n h)(h' - \Pi_n h')\right| \leq 2P\partial \ell^2_{\mu_0} |\Pi_0 h - \Pi_n h| \tag{A.13}$$

using the fact that $(h' - \Pi_n h') \in \mathcal{H}(2)$. By the triangle inequality

$$P\partial\ell^2_{\mu_0} |\Pi_0 h - \Pi_n h| \leq P\partial\ell^2_{\mu_0} |\Pi_0 h - \Pi_\rho h| + P\partial\ell^2_{\mu_0} |\Pi_\rho h - \Pi_n h| .$$

By Lemma 11, the r.h.s. is less than a constant multiple of $\sqrt{P\partial^2\ell_{\mu_0} (\Pi_0 h - \Pi_\rho h)^2}$. By Lemma 9, this goes to zero as $\rho \to 0$. By Holder inequality and then Lemma 8, the second term on the r.h.s. of the above display is $o_p(1)$ when $\rho = \rho_n$ satisfy the condition of that lemma. These remarks imply that (A.13) is $o_p(1)$. The last term in (A.12) is bounded similarly. The uniform convergence of the covariance is proved because all the terms in (A.12) converge to zero uniformly in $h, h' \in \mathcal{H}^K(1)$.

Now, we show Point 2. Recall that we have defined eigenvalues and eigenvectors relative to the standard inner product divided by the number of elements R. This is itself an inner product. In particular, note that in this case the eigenvectors and eigenvalues are $R^{1/2}$ and R^{-1} times the usual matrix eigenvectors and eigenvalues, respectively. Then, Point 2. follows from the inequality

$$\sup_{k>0} |\omega_{nk} - \omega_k| \leq \frac{1}{R} \sum_{h \in \tilde{\mathcal{R}}_1} |\Sigma_n(h,h) - \Sigma(h,h)| ,$$

which uses [6, Lemma 4.2] together with the fact that the operator norm of a covariance function is bounded by the nuclear norm [6]. Clearly, the r.h.s. is bounded by $\sup_{h \in \mathcal{H}^K(1)} |\Sigma_n(h,h) - \Sigma(h,h)|$ which converges to zero in probability.

Finally we show Point 3. By definition of the eigenvalues and eigenvectors, $\Sigma(h,h) = \sum_{k=1}^\infty \omega_k \psi_k(h) \psi_k(h)$ so that

$$\frac{1}{R} \sum_{h \in \tilde{\mathcal{R}}_1} \Sigma(h,h) = \sum_{k=1}^\infty \omega_k \leq \sup_{h \in \mathcal{H}^K(1)} \Sigma(h,h) < \infty$$

implying that the eigenvalues are summable. The sum of the sample eigenvalues is equal to

$$\begin{aligned} \frac{1}{R} \sum_{h \in \tilde{\mathcal{R}}_1} \Sigma_n(h,h) &\leq \frac{1}{R} \sum_{h \in \tilde{\mathcal{R}}_1} \Sigma(h,h) + \frac{1}{R} \sum_{h \in \tilde{\mathcal{R}}_1} |\Sigma_n(h,h) - \Sigma(h,h)| \\ &\leq \sup_{h \in \mathcal{H}^K(1)} \Sigma(h,h) + \sup_{h \in \mathcal{H}^K(1)} |\Sigma_n(h,h) - \Sigma(h,h)| . \end{aligned}$$

As shown above, the first term on the r.h.s. is finite and the second term converges to zero in probability. Hence, the sample eigenvalues are summable in probability. In particular, from these remarks we deduce that for any $c < \infty$ such that $c \geq 1 + \sum_{k=1}^\infty \omega_k$, we have that $\Pr(\sum_{k=1}^\infty \omega_{n,k} > c) = o(1)$. □

A.5. Proof of Theorem 1

To keep the notation more compact, we introduce additional concepts to be used in the rest of the proofs. This will also make reference to existing results

easier. Let $l^\infty\left(\mathcal{H}^K\right)$ be the space of uniformly bounded functions on $\mathcal{H}^K$. Let $\Psi(\mu)$ be the operator in $l^\infty\left(\mathcal{H}^K\right)$ such that $\Psi(\mu)h = P\partial\ell_\mu h$, $h \in \mathcal{H}^K$. When $\mu_0 \in \text{int}\left(\mathcal{H}^K(B)\right)$, it holds that $\Psi(\mu_0)h = 0$, for any $h \in \mathcal{H}^K(1)$. The empirical counterpart of $\Psi(\mu)$ is the operator $\Psi_n(\mu)$ such that $\Psi_n(\mu)h = P_n\partial\ell_\mu h$. Finally, write $\dot{\Psi}_{\mu_0}(\mu - \mu_0)$ for the Fréchet derivative of $\Psi(\mu)$ at μ_0 tangentially to $(\mu - \mu_0)$, where $\mu, \mu_0 \in \mathcal{H}^K(B)$. Then, $\dot{\Psi}_{\mu_0}$ is an operator from $\mathcal{H}^K$ to $l^\infty\left(\mathcal{H}^K\right)$. This same notation is used in [31, Chapter 3.3].

Given that $|P_x\Delta_3|_\infty < \infty$ and $|\mu_{0n} - \mu_0|_2 = o_P\left(n^{1/4}\right)$, Equation (26) in [23] holds, and we have that

$$\sqrt{n}\Psi_n(\mu_{0n}) = \sqrt{n}\Psi_n(\mu_0) + \dot{\Psi}_{\mu_0}\sqrt{n}(\mu_{0n} - \mu_0) + o_p(1). \tag{A.14}$$

Trivially, any $h \in \mathcal{H}^K(1)$ can be written as $h = \Pi_\rho h + (h - \Pi_\rho h)$. By Lemma 5, $(h - \Pi_\rho h) \in \mathcal{H}^K(2)$. Then, $\sqrt{n}\Psi_n(\mu_0)(h - \Pi_\rho h)$ for $h \in \mathcal{H}^K(1)$ is mean zero and converges weakly to a Gaussian process with a.s. continuous sample paths by the Donsker Theorem ([31, Theorem 2.5.6]) because by Lemma 1 the entropy integral is finite (see also [23, Lemma 5]. Therefore, (A.14) also applies to $\Psi_n(\mu)$ as an element in the space of uniformly bounded functions on $\mathcal{H}^K(2)$. Now, for $\rho = \rho_n$,

$$\begin{aligned}\sqrt{n}\Psi_n(\mu_{0n})(h - \Pi_{n,\rho}h) =& \sqrt{n}\Psi_n(\mu_{0n})(h - \Pi_\rho h) \\ &+ \sqrt{n}\Psi_n(\mu_{0n})(\Pi_\rho - \Pi_{n,\rho})h\end{aligned} \tag{A.15}$$

adding and subtracting $\sqrt{n}\Psi_n(\mu_{0n})\Pi_\rho h$. By definition, the operator $\Psi_n(\mu_{0n})$ is such that

$$\sqrt{n}\Psi_n(\mu_{0n})(\Pi_\rho - \Pi_{n,\rho})h = \sqrt{n}P_n\partial\ell_{\mu_{0n}}(\Pi_\rho - \Pi_{n,\rho})h.$$

Hence, using the second part of Lemma 8, the r.h.s. of (A.15) is equal to

$$\sqrt{n}\Psi_n(\mu_{0n})(h - \Pi_\rho h) + o_p(1).$$

By (A.14), this is in turn equal to

$$\sqrt{n}\Psi_n(\mu_0)(h - \Pi_\rho h) + \sqrt{n}\dot{\Psi}_{\mu_0}(\mu_{0n} - \mu_0)(h - \Pi_\rho h) + o_p(1).$$

Using linearity, rewrite

$$\begin{aligned}\dot{\Psi}_{\mu_0}\sqrt{n}(\mu_{0n} - \mu_0)(h - \Pi_\rho h) =& \dot{\Psi}_{\mu_0}\sqrt{n}(\mu_{0n} - \mu_0)(h - \Pi_0 h) \\ &+ \dot{\Psi}_{\mu_0}\sqrt{n}(\mu_{0n} - \mu_0)(\Pi_0 h - \Pi_\rho h).\end{aligned}$$

The first term on the r.h.s. is $P\partial^2\ell_{\mu_0}\sqrt{n}(\mu_{0n} - \mu_0)(h - \Pi_0 h)$. This is zero because $(\mu_{0n} - \mu_0)$ is in the linear span of elements in $\mathcal{R}_0$, and $(h - \Pi_0)$ is orthogonal to any element in $\mathcal{R}_0$ (w.r.t. $\tilde{P}$ by (3.1) with $\rho = 0$). Using Holder inequality, Lemma A.10 shows that the absolute value of the second term on the r.h.s. of the display is $o_p(1)$.

We deduce that the asymptotic distribution of $\sqrt{n}\Psi_n(\mu_{0n})(h - \Pi_n h)$ is given by the one of $\sqrt{n}\Psi_n(\mu_0)(h - \Pi_\rho h)$ for $\rho \to 0$ at a suitable rate. By Lemma 12 and the definition of $\Psi_n(\mu_0)$, the latter converges weakly to a centered Gaussian process as in the statement of Theorem 1.

A.6. Proof of Proposition 2

This follows from Point 1. in Lemma 13.

A.7. Proof of Proposition 3

Note that $\left\{G\left(h-\Pi_0 h\right): h\in\tilde{\mathcal{R}}_1\right\}$ is an R-dimensional Gaussian vector. Hence, by the spectral decomposition, the distribution of its squared Euclidean norm is given by the weighted sum of independent squared standard normal random variables scaled by the eigenvalues of the covariance matrix. Given that we are scaling by the number of elements R in the vector, the eigenvectors $\{\psi_k(h): h\in\mathcal{R}_1\}$ are $R^{1/2}$ times the standard matrix eigenvectors, while the eigenvalues $\{\omega_k: k\geq 1\}$ are R^{-1} times the standard matrix eigenvalues. Hence the result follows as (3.5) is the squared Euclidean norm divided by R.

A.8. Proof of Theorem 2

The test statistic $\hat{S}_n$ is the square of $\sqrt{n}\Psi_n\left(\mu_{0n}\right)\left(h-\Pi_{n,\rho}h\right)$ averaged over a finite number of functions h. By Theorem 1 and the continuous mapping theorem its distribution is given by S in (3.5). The latter admits a series representation as given in Proposition 3. The distribution of the approximation to S when the sample eigenvalues are used is $\hat{S}:=\sum_{k\geq 1}\omega_{nk}N_k^2$. By the triangle inequality,

$$\left|\hat{S}-S\right|\leq\sum_{k=1}^{\infty}\left|\omega_{nk}-\omega_k\right|N_k^2. \tag{A.16}$$

The sum can be split into two parts, one for $k\leq L$ plus one for $k>L$ where here L is a positive integer. Hence, deduce that the above is bounded by

$$L\sup_{k\leq L}\left|\omega_{nk}-\omega_k\right|N_k^2+\sum_{k>L}\left(\omega_{nk}+\omega_k\right)N_k^2$$

Using Lemma 13, the first term is $o_p(1)$ for any fixed integer L. By Lemma 13, again, there is a positive summable sequence $(a_k)_{k\geq 1}$ such that, as $n\to\infty$, the event $\left\{\sup_{k\geq 1}\omega_{nk}a_k^{-1}=\infty\right\}$ is contained in the event $\{\sum_{k=1}^{\infty}\omega_{n,k}>c\}$ for some finite constant c. However, by Lemma 13, the latter event has probability going to zero for finite c as given in that lemma. Hence, the second term in the display is bounded with probability going to one by

$$\left(\sup_{k>0}\omega_{nk}a_k^{-1}\right)\sum_{k>L}a_kN_k^2+\sum_{k>L}\omega_kN_k^2,$$

where $\sup_{k>0}\omega_{nk}a_k^{-1}=O_p(1)$. Given that

$$\mathbb{E}\left[\sum_{k>L}a_kN_k^2+\sum_{k>L}\omega_kN_k^2\right]\leq\sum_{k>L}\left(a_k+\omega_k\right)\to 0$$

as $L \to \infty$, deduce that letting $L \to \infty$ slowly enough, (A.16) is $o_p(1)$. Hence, we have shown that both $\hat{S}_n$ and $\hat{S}$ converge in distribution to S. In fact, $\hat{S}$ converges in probability.

A.9. Proof of Corollary 1

This is a consequence of the last statement in Lemma 8.

Appendix B: Additional numerical details

Tables 7–16 report more simulation results. The column heading "No Π" means that no correction was used in estimating the test statistic and the covariance function. This means that instead of using $(h - \Pi_n h)$ we just use h, which is the naive estimator in the presence of a nuisance parameter. The column heading "Size" stands for the nominal size and the simulated frequency of rejection should be close to this when the null is true. For the short names used in the tables, refer to Tables 1, 2, and 4. The model Lin1Poly2 used in the infinite dimensional model is defined in (4.1). The model LinPoly2Local is defined in (4.2).

TABLE 7

Finite Dimensional Model. Simulated frequency of rejections for $n = 100$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The true model is Lin3. Restrictions Lin1 and Lin2 should be rejected. The column heading "Size" stands for the nominal size.

			Lin1		Lin2		Lin3		LinAll	
ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π	No Π	Π	No Π	Π
0	1	0.10	1.00	1.00	0.99	0.99	0.08	0.12	0.05	0.13
0	1	0.05	1.00	1.00	0.96	0.98	0.03	0.06	0.02	0.07
0	0.2	0.10	0.71	0.78	0.44	0.50	0.08	0.12	0.05	0.13
0	0.2	0.05	0.54	0.66	0.25	0.36	0.04	0.06	0.02	0.07
0.75	1	0.10	0.91	0.95	0.21	0.31	0.05	0.10	0.07	0.14
0.75	1	0.05	0.80	0.90	0.12	0.20	0.02	0.05	0.03	0.07
0.75	0.2	0.10	0.28	0.39	0.08	0.14	0.05	0.10	0.07	0.14
0.75	0.2	0.05	0.16	0.25	0.03	0.06	0.02	0.05	0.03	0.07

TABLE 8

Finite Dimensional Model. Simulated frequency of rejections for $n = 1000$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The true model is Lin3. Restrictions Lin1 and Lin2 should be rejected. The column heading "Size" stands for the nominal size.

			Lin1		Lin2		Lin3		LinAll	
ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π	No Π	Π	No Π	Π
0	1	0.10	1.00	1.00	1.00	1.00	0.09	0.11	0.07	0.10
0	1	0.05	1.00	1.00	1.00	1.00	0.05	0.05	0.04	0.06
0	0.2	0.10	1.00	1.00	1.00	1.00	0.10	0.11	0.07	0.11
0	0.2	0.05	1.00	1.00	1.00	1.00	0.05	0.05	0.04	0.06
0.75	1	0.10	1.00	1.00	1.00	1.00	0.06	0.09	0.06	0.10
0.75	1	0.05	1.00	1.00	1.00	1.00	0.03	0.05	0.03	0.04
0.75	0.2	0.10	1.00	1.00	0.48	0.60	0.06	0.09	0.06	0.10
0.75	0.2	0.05	1.00	1.00	0.32	0.45	0.03	0.05	0.03	0.04

TABLE 9

Finite Dimensional Model. Simulated frequency of rejections for $n = 100$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The true model is LinAll. Restrictions Lin1, Lin2, and Lin3 should be rejected. The column heading "Size" stands for the nominal size.

			Lin1		Lin2		Lin3		LinAll	
ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π	No Π	Π	No Π	Π
0	1	0.10	1.00	1.00	1.00	1.00	1.00	1.00	0.06	0.11
0	1	0.05	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.05
0	0.2	0.10	0.84	0.88	0.80	0.85	0.77	0.82	0.05	0.13
0	0.2	0.05	0.71	0.77	0.65	0.75	0.62	0.72	0.02	0.07
0.75	1	0.10	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.14
0.75	1	0.05	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.07
0.75	0.2	0.10	0.95	0.97	0.89	0.93	0.80	0.87	0.07	0.14
0.75	0.2	0.05	0.89	0.94	0.80	0.89	0.66	0.80	0.03	0.07

TABLE 10

Finite Dimensional Model. Simulated frequency of rejections for $n = 1000$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The true model is LinAll. Restrictions Lin1, Lin2, and Lin3 should be rejected. The column heading "Size" stands for the nominal size.

			Lin1		Lin2		Lin3		LinAll	
ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π	No Π	Π	No Π	Π
0	1	0.10	1.00	1.00	1.00	1.00	1.00	1.00	0.49	0.08
0	1	0.05	1.00	1.00	1.00	1.00	1.00	1.00	0.23	0.05
0	0.2	0.10	1.00	1.00	1.00	1.00	1.00	1.00	0.08	0.10
0	0.2	0.05	1.00	1.00	1.00	1.00	1.00	1.00	0.04	0.06
0.75	1	0.10	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.10
0.75	1	0.05	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.05
0.75	0.2	0.10	1.00	1.00	1.00	1.00	1.00	1.00	0.06	0.10
0.75	0.2	0.05	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.05

TABLE 11

Finite Dimensional Model. Simulated frequency of rejections for $n = 100$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The the true model is Lin1Poly4. Restrictions Lin1, Lin2, Lin3, and LinAll should be rejected. The column heading "Size" stands for the nominal size.

			Lin1		Lin2		Lin3		LinAll		Lin1Poly	
ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π	No Π	Π	No Π	Π	No Π	Π
0	1	0.10	0.97	0.94	0.97	0.94	0.97	0.95	0.59	0.61	0.03	0.15
0	1	0.05	0.94	0.91	0.95	0.92	0.95	0.92	0.54	0.50	0.01	0.09
0	0.2	0.10	0.71	0.72	0.72	0.74	0.73	0.75	0.30	0.31	0.04	0.12
0	0.2	0.05	0.57	0.62	0.58	0.64	0.61	0.68	0.23	0.23	0.01	0.06
0.75	1	0.10	0.94	0.95	0.93	0.93	0.85	0.88	0.61	0.61	0.02	0.13
0.75	1	0.05	0.89	0.92	0.86	0.90	0.69	0.79	0.54	0.52	0.01	0.06
0.75	0.2	0.10	0.70	0.77	0.57	0.66	0.32	0.39	0.33	0.31	0.01	0.14
0.75	0.2	0.05	0.55	0.68	0.39	0.53	0.17	0.28	0.25	0.24	0.00	0.08

TABLE 12

Finite Dimensional Model. Simulated frequency of rejections for $n = 1000$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The true model is Lin1Poly4. Restrictions Lin1, Lin2, Lin3, and LinAll should be rejected. The column heading "Size" stands for the nominal size.

			Lin1		Lin2		Lin3		LinAll		Lin1Poly	
ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π	No Π	Π	No Π	Π	No Π	Π
0	1	0.10	1	1	1	1	1	1	0.92	0.91	0.03	0.1
0	1	0.05	1	1	1	1	1	1	0.9	0.88	0.02	0.05
0	0.2	0.10	1	0.99	1	0.99	1	0.99	0.75	0.72	0.04	0.1
0	0.2	0.05	1	0.98	1	0.98	1	0.99	0.7	0.65	0.01	0.05
0.75	1	0.10	1	1	1	1	1	1	0.89	0.87	0.02	0.09
0.75	1	0.05	1	1	1	1	1	1	0.87	0.84	0	0.05
0.75	0.2	0.10	1	0.99	1	0.99	0.99	0.99	0.75	0.73	0.02	0.11
0.75	0.2	0.05	1	0.99	1	0.99	0.99	0.98	0.72	0.67	0.01	0.06

TABLE 13

Infinite Dimensional Model. Simulated frequency of rejections for $n = 100$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The true model is Lin1Poly2. Restriction LinAll should be rejected. The column heading "Size" stands for the nominal size.

			Lin1NonLin2		LinAll	
ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π
0	1	0.10	0.00	0.12	0.00	0.91
0	1	0.05	0.00	0.06	0.00	0.84
0	0.2	0.10	0.00	0.12	0.00	0.42
0	0.2	0.05	0.00	0.06	0.00	0.31
0.75	1	0.10	0.00	0.11	0.00	0.97
0.75	1	0.05	0.00	0.06	0.00	0.93
0.75	0.2	0.10	0.00	0.11	0.00	0.51
0.75	0.2	0.05	0.00	0.06	0.00	0.38

TABLE 14

Infinite Dimensional Model. Simulated frequency of rejections for $n = 1000$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The true model is Lin1Poly2. Restriction LinAll should be rejected. The column heading "Size" stands for the nominal size.

			Lin1NonLin2		LinAll	
ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π
0	1	0.10	0.00	0.09	0.99	1.00
0	1	0.05	0.00	0.04	0.82	1.00
0	0.2	0.10	0.00	0.09	0.00	1.00
0	0.2	0.05	0.00	0.04	0.00	1.00
0.75	1	0.10	0.00	0.11	1.00	1.00
0.75	1	0.05	0.00	0.03	1.00	1.00
0.75	0.2	0.10	0.00	0.11	0.17	1.00
0.75	0.2	0.05	0.00	0.03	0.02	1.00

TABLE 15

Infinite Dimensional Model. Simulated frequency of rejections for $n = 5000$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The true model is Lin1Poly2. Restriction LinAll should be rejected. The column heading "Size" stands for the nominal size.

			Lin1NonLin2		LinAll	
ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π
0	1	0.10	0	0.12	1	1
0	1	0.05	0	0.05	1	1
0	0.2	0.10	0	0.12	1	1
0	0.2	0.05	0	0.05	0.97	1
0.75	1	0.10	0	0.11	1	1
0.75	1	0.05	0	0.06	1	1
0.75	0.2	0.10	0	0.11	1	1
0.75	0.2	0.05	0	0.06	1	1

TABLE 16

Infinite Dimensional Model. Simulated frequency of rejections for $n = 100$, and various combinations of signal to noise ratio $\sigma^2_{\mu/\varepsilon}$, and variables correlation ϱ. The true model is Lin1Poly2Local. Restriction LinAll should be rejected. The column heading "Size" stands for the nominal size. The local power is reported for different deviations from the lower dimensional model via the constant c defined after (4.2).

ϱ	$\sigma^2_{\mu/\varepsilon}$	Size	No Π	Π	No Π	Π	No Π	Π
					c			
			0.1		0.5		1	
0	1	0.10	0.00	0.10	0.00	0.13	0.00	0.18
0	1	0.05	0.00	0.06	0.00	0.06	0.00	0.10
0	0.2	0.10	0.00	0.10	0.00	0.11	0.00	0.12
0	0.2	0.05	0.00	0.06	0.00	0.05	0.00	0.06
0.75	1	0.10	0.00	0.12	0.00	0.14	0.00	0.21
0.75	1	0.05	0.00	0.06	0.00	0.07	0.00	0.13
0.75	0.2	0.10	0.00	0.12	0.00	0.12	0.00	0.13
0.75	0.2	0.05	0.00	0.07	0.00	0.06	0.00	0.07
					c			
			2		5		10	
0	1	0.10	0.00	0.38	0.00	0.93	0.01	1.00
0	1	0.05	0.00	0.27	0.00	0.86	0.00	0.99
0	0.2	0.10	0.00	0.16	0.00	0.46	0.00	0.88
0	0.2	0.05	0.00	0.09	0.00	0.32	0.00	0.79
0.75	1	0.10	0.00	0.49	0.00	0.98	0.26	1.00
0.75	1	0.05	0.00	0.35	0.00	0.95	0.06	1.00
0.75	0.2	0.10	0.00	0.19	0.00	0.56	0.00	0.95
0.75	0.2	0.05	0.00	0.12	0.00	0.43	0.00	0.91

Acknowledgments

I am grateful to the Editor Domenico Marinucci and the Referees for comments that have lead to substantial improvements both in content and presentation.

References

[1] Adler, R.J. and J.E. Taylor (2007) Random Fields and Geometry. New York: Springer. MR2319516

[2] Andrews, D.W.K. (1994) Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity. Econometrica 62, 43–72. MR1258665
[3] Banerjee, A., D. Dunson and S. Todkar (2008) Efficient Gaussian Process Regression for Large Data Sets. Biometrika 94, 1-16. MR3034325
[4] Belloni, A., V. Chernozhukov, I. Hansen (2017) Program Evaluation and Causal Inference With High-Dimensional Data. Econometrica 85, 233–298. MR3611771
[5] Bierens, H.J. (1982) Consistent Model Specification Tests. Journal of Econometrics 20, 105–134. MR0685673
[6] Bosq, D. (2000) Liner Processes in Function Spaces. Berlin: Springer. MR1783138
[7] Chen, X. and Z. Liao (2014) Sieve M Inference on Irregular Parameters. Journal of Econometrics 182, 70–86. MR3212762
[8] Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018) Double/Debiased Machine Learning for Treatment and Structural Parameters. The Econometrics Journal 21, 1–68. MR3769544
[9] Christmann, A. and I. Steinwart (2007) Consistency and Robustness of Kernel-Based Regression in Convex Risk Minimization. Bernoulli 13, 799–719. MR2348751
[10] Connor, G., M. Hagmann and O. Linton (2012) Efficient Semiparametric Estimation of the Fama-French Model and Extensions. Econometrica 80, 713–754. MR2951947
[11] Delgado, M.A. and W.G. Manteiga (2001) Significance Testing in Nonparametric Regression Based on the Bootstrap. The Annals of Statistics 29, 1469–1507. MR1873339
[12] Escanciano, J.C., D.T. Jacho-Chávez and A. Lewbel (2014) Uniform Convergence of Weighted Sums of Non and Semiparametric Residuals for Estimation and Testing. Journal of Econometrics 178, 426–443. MR3132442
[13] Fan, J., C.M. Zhang, and J. Zhang (2001) Generalized Likelihood Ratio Statistics and Wilks Phenomenon. The Annals of Statistics 29, 153–193. MR1833962
[14] Fan, J. and J. Jiang (2005) Nonparametric Inferences for Additive Models. Journal of the American Statistical Association 100, 890–907. MR2201017
[15] Hable, R. (2012) Asymptotic Normality of Support Vector Machine Variants and Other Regularized Kernel Methods. Journal of Multivariate Analysis 106, 92–117. MR2887683
[16] Härdle, W. and E. Mammen (1993) Comparing Nonparametric Versus Parametric Regression Fits. Annals of Statistics 21, 1926–1947. MR1245774
[17] Hartigan, J.A. (2014) Bounding the Maximum of Dependent Random Variables. Electronic Journal of Statistics 8, 3126–3140. MR3301303
[18] Hudson, A., M. Carone and A. Shojaie (2021) Inference on Function-Valued Parameters Using a Restricted Score Test. https://arxiv.org/abs/2105.06646.
[19] Lázaro-Gredilla, M., J. Quiñonero-Candela, C.E. Rasmussen and A.R.

Figueiras-Vidal (2010) Sparse Spectrum Gaussian Process Regression. Journal of Machine Learning Research 11, 1865–1881. MR2660655
[20] Li, W.V. and W. Linde (1999) Approximation, Metric Entropy and Small Ball Estimates for Gaussian Measures. Annals of Probability 27, 1556–1578. MR1733160
[21] Rasmussen, C. and C.K.I. Williams (2006) Gaussian Processes of Machine Learning. Cambridge, MA: MIT Press. MR2514435
[22] Ritter, K., G.W. Wasilkowski and H. Wozniakowski (1995) Multivariate Integration and Approximation for Random Fields Satisfying Sacks-Ylvisaker Conditions. Annals of Applied Probability 5, 518–540. MR1336881
[23] Sancetta A. (2021) Estimation in Reproducing Kernel Hilbert Spaces with Dependent Data. IEEE Transactions on Information Systems 67, 1782–1795. MR4282326
[24] Schölkopf, B., R. Herbrich, and A.J. Smola (2001) A Generalized Representer Theorem. In D. Helmbold and B. Williamson (eds.) Neural Networks and Computational Learning Theory 81, 416–426. Berlin: Springer. MR2042050
[25] Shen, X. and J. Shi (2005) Sieve Likelihood Ratio Inference on General Parameter Space. Science in China Series A: Mathematics 48, 67–78. MR2156616
[26] Steinwart, I., and A. Christmann (2008) Support Vector Machines. Berlin: Springer. MR2796580
[27] Stute, W. (1997) Nonparametric Model Checks for Regression. The Annals of Statistics 25, 613–641. MR1439316
[28] Stute, W., W. González Manteiga and M. Presedo Quindimil (1998a) Bootstrap Approximations in Model Checks for Regression. Journal of the American Statistical Association 93, 141–149. MR1614600
[29] Stute, W., S. Thies and L.-X. Zhu (1998b) Model Checks for Regression: An Innovation Process Approach. The Annals of Statistics 26, 1916–1934. MR1673284
[30] van der Vaart, A. (1998) Asymptotic Statistics. Cambridge: Cambridge University Press. MR1652247
[31] van der Vaart, A. and J.A. Wellner (2000) Weak Convergence and Empirical Processes. New York: Springer. MR1385671
[32] van der Vaart, A. and J.H. van Zanten (2008) Reproducing Kernel Hilbert Spaces of Gaussian Priors. Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, 200–222. Beachwood, Ohio: Institute of Mathematical Statistics. MR2459226
[33] Wahba, G. (1990) Spline Models for Observational Data. Philadelphia: SIAM. MR1045442
[34] Wooldridge, J.M. (1990) A Unified Approach to Robust, Regression-Based Specification Tests. Econometric Theory 6, 17–43. MR1059144
[35] Wooldridge, J.M. (2007) Inverse Probability Weighted Estimation for General Missing Data Problems. Journal of Econometrics 141, 1281–1301. MR2413502
[36] Tsybakov, A.B. (2009) Introduction to Nonparametric Estimation. New York: Springer. MR2724359