Received: 26 January 2021 | Accepted: 27 May 2022 | Published on: 2 November 2022

DOI: 10.1111/rssb.12544

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Series B Statistical Methodology

ORIGINAL ARTICLE

Conditional independence testing in Hilbert spaces with applications to functional data analysis

Anton Rask Lundborg[1,*] | **Rajen D. Shah**[1,†] | **Jonas Peters**[2,‡]

[1]University of Cambridge, Cambridge, UK

[2]University of Copenhagen, Copenhagen, Denmark

Correspondence
Rajen D. Shah, University of Cambridge, Cambridge, UK.
Email: r.shah@statslab.cam.ac.uk

Abstract

We study the problem of testing the null hypothesis that X and Y are conditionally independent given Z, where each of X, Y and Z may be functional random variables. This generalises testing the significance of X in a regression model of scalar response Y on functional regressors X and Z. We show, however, that even in the idealised setting where additionally (X, Y, Z) has a Gaussian distribution, the power of any test cannot exceed its size. Further modelling assumptions are needed and we argue that a convenient way of specifying these assumptions is based on choosing methods for regressing each of X and Y on Z. We propose a test statistic involving inner products of the resulting residuals that is simple to compute and calibrate: type I error is controlled uniformly when the in-sample prediction errors are sufficiently small. We show this requirement is met by ridge regression in functional linear model settings without requiring any eigen-spacing conditions or lower bounds on the eigenvalues of the covariance of the functional regressor. We apply our test in constructing confidence intervals

*Part of this work was done while ARL was at the University of Copenhagen. ARL was supported by the Cantab Capital Institute for the Mathematics of Information.

[†]RDS was supported by an EPSRC Programme Grant EP/N031938/1 and an EPSRC First Grant EP/R013381/1.

[‡]JP was supported by the Carlsberg Foundation and a research grant (18968) from the VILLUM Foundation.

for truncation points in truncated functional linear models and testing for edges in a functional graphical model for EEG data.

KEYWORDS

functional graphical model, function-on-function regression, significance testing, truncated functional linear model, uniform type I error control

1 | INTRODUCTION

In a variety of application areas, such as meteorology, neuroscience, linguistics and chemometrics, we observe samples containing random functions (Ramsay & Silverman, 2005; Ullah & Finch, 2013). The field of functional data analysis (FDA) has a rich toolbox of methods for the study of such data. For instance, there are a number of regression methods for different functional data types, including linear function-on-scalar (Reiss et al., 2010), scalar-on-function (Delaigle & Hall, 2012; Goldsmith et al., 2011; Hall & Horowitz, 2007; Reiss & Ogden, 2007; Shin 2009; Yuan & Cai, 2010) and function-on-function (Ivanescu et al., 2015; Scheipl et al., 2015) regression; there are also nonlinear and nonparametric variants (Fan et al., 2015; Ferraty & Vieu, 2006; Ferraty et al., 2011; Yao & Müller, 2010) and versions able to handle potentially large numbers of functional predictors (Fan et al., 2015) to give a few examples; see Wang et al. (2016), Morris (2015) for helpful reviews and a more extensive list of relevant references. The availability of software packages for functional regression methods, such as the R-packages `refund` (Goldsmith et al., 2020) and `FDboost` (Brockhaus et al., 2020), allow practitioners to easily adopt the FDA framework for their particular data.

One area of FDA that has received less attention is that of conditional independence testing. Given random elements X, Y, Z, the conditional independence $X \perp\!\!\!\perp Y \mid Z$ formalises the idea that X contains no further information about Y beyond that already contained in Z. A precise definition is given in Section 1.2. Inferring conditional independence from observed data is of central importance in causal inference (Pearl, 2009; Peters et al., 2017; Spirtes et al. 2000), graphical modelling (Koller & Friedman, 2009; Lauritzen, 1996) and variable selection. For example, consider the linear scalar-on-function regression model

$$Y = \int_0^1 \theta_X(t)X(t)dt + \int_0^1 \theta_Z(t)Z(t)dt + \varepsilon, \tag{1}$$

where X, Z are random covariate functions taking values in $L^2([0,1],\mathbb{R})$, θ_X, θ_Z are unknown parameter functions, $Y \in \mathbb{R}$ is a scalar response and $\varepsilon \in \mathbb{R}$ satisfying $\varepsilon \perp\!\!\!\perp (X, Z)$ represents stochastic error. In this model, conditional independence $X \perp\!\!\!\perp Y \mid Z$ is equivalent to $\theta_X = 0$, that is, whether the functional predictor X is significant.

For nonlinear regression models, the conditional independence $X \perp\!\!\!\perp Y \mid Z$ still characterises whether X is useful for predicting Y given Z. Indeed, consider a more general setting where Y is a potentially infinite-dimensional response, and $X_1, \dots, X_p$ are predictors, some or all of which may be functional. Then a set of predictors $S \subseteq \{1, \dots, p\}$ that contain all useful information for predicting Y, that is such that $Y \perp\!\!\!\perp \{X_j\}_{j \notin S} \mid \{X_j\}_{j \in S}$, is known as a Markov blanket of Y in the graphical modelling literature (Pearl, 2014; Section 3.2.1). If $Y \not\perp\!\!\!\perp X_j \mid \{X_k\}_{k \neq j}$, then j is contained in every Markov blanket, and under mild conditions (e.g. the intersection property Pearl (2009), Peters (2014)), the smallest Markov blanket (sometimes called the Markov boundary) is unique and coincides exactly with those variables j satisfying this conditional dependence. This set may thus be inferred by applying conditional independence tests. Conditional independence tests may also be used to test for edge presence in conditional independence graphs and are at the heart of several methods for causal discovery (Peters et al., 2016; Spirtes et al., 2000).

Recent work (Shah & Peters, 2020), however, has shown that in the setting where X,Y and Z are random vectors where Z is absolutely continuous (i.e. has a density with respect to Lebesgue measure), testing the conditional independence $X \perp\!\!\!\perp Y \mid Z$ is fundamentally hard in the sense that any test for conditional independence must have power at most its size. Intuitively, the reason for this is that given any test, there are potentially highly complex joint distributions for the triple (X, Y, Z) that maintain conditional independence but yield rejection rates as high as for any alternative distribution. Lipschitz constraints on the joint density, for example, preclude the presence of such distributions (Neykov et al., 2020).

In the context of functional data, however, the problem can be more severe, and we show in this work that even in the idealised setting where (X, Y, Z) are jointly Gaussian in the functional linear regression model (1), testing for $X \perp\!\!\!\perp Y \mid Z$ is fundamentally impossible: any test must have power at most its size. In other words, any test with power β at some alternative cannot hope to control type I error at level $\alpha < \beta$ across the entirety of the null hypothesis, even if we are willing to assume Gaussianity. Perhaps more surprisingly, this fundamental problem persists even if additionally we allow ourselves to know the precise null distribution of the infinite-dimensional Z.

Consequently, there is no general purpose conditional independence test even for Gaussian functional data, and we must necessarily make some additional modelling assumptions to proceed. We argue that this calls for the need of conditional independence tests whose suitability for any functional data setting can be judged more easily.

Motivated by the Generalised Covariance Measure (Shah & Peters, 2020), we propose a simple test we call the Generalised Hilbertian Covariance Measure (GHCM) that involves regressing X on Z and Y on Z (each of which may be functional or indeed collections of functions), and computing a test statistic formed from inner products of pairs of residuals. We show that the validity of this form of test relies primarily on the relatively weak requirement that the regression procedures have sufficiently small in-sample prediction errors. We thus aim to convert the problem of conditional independence testing into the more familiar task of regression with functional data, for which well-developed methods are readily available. These features mark out our test as rather different from existing approaches for assessing conditional independence in FDA, which we review in the following.

One approach to measuring conditional dependence with functional data is based on the Gaussian graphical model. Zhu et al. (2016) propose a Bayesian approach for learning a graphical model for jointly Gaussian multivariate functional data. Qiao et al. (2019) and Zapata et al. (2019) study approaches based on generalisations of the graphical Lasso (Yuan & Lin, 2007). These latter methods do not aim to perform statistical tests for conditional independence, but rather provide a point estimate of the graph, for which the authors establish consistency results valid in potentially high-dimensional settings.

As discussed earlier, conditional independence testing is related to significance testing in regression models. There is, however, a paucity of literature on formal significance tests for functional predictors. The R implementation (Goldsmith et al., 2020) of the popular functional regression methodology of Greven and Scheipl (2017) produces p-values for the inclusion of a functional predictor based on significance tests for generalised additive models developed in Wood (2013). These tests, while being computationally efficient, however, do not have formal uniform level control guarantees.

1.1 | Our main contributions and organisation of the paper

1. **It is impossible to test conditional independence with Gaussian functional data.** In Section 2 we present our formal hardness result on conditional independence testing for Gaussian functional data. The proof rests on a new result on the maximum power attainable at any alternative when testing for conditional independence with multivariate Gaussian data. The full technical details are given in Section A of the supplementary material. As we cannot hope to have level control uniformly over the entirety of the null of conditional independence, it is important to establish, for any given test, subsets $\tilde{\mathcal{P}}_0$ of null distributions $\mathcal{P}_0$ over which we do have uniform level control.
2. **We provide new tools allowing for the development of uniform results in FDA.** Uniform results are scarce in functional data analysis; we develop the tools for deriving such results in Section B of the supplementary material which studies uniform convergence of Hilbertian and Banachian random variables.
3. **Given sufficiently good methods for regressing each of X and Y on Z, the GHCM can test conditional independence with certain uniform level guarantees.** In Section 3 we describe our new GHCM testing framework for testing $X \perp\!\!\!\perp Y \mid Z$, where each of X, Y and Z may be collections of functional and scalar variables. In Section 4 we show that for the GHCM, an effective null hypothesis $\tilde{\mathcal{P}}_0$ may be characterised as one where in addition to some tightness and moment conditions, the conditional expectations $\mathbb{E}(X \mid Z)$ and $\mathbb{E}(Y \mid Z)$ can be estimated at sufficiently fast rates, such that the product of the corresponding in-sample mean squared prediction errors (MSPEs) decay faster than $1/n$ uniformly, where n is the sample size. Note that this does not contradict the hardness result: it is well known that there do not exist regression methods with risk converging to zero uniformly over all distributions for the data (Györfi et al., 2002, Theorem 3.1). Thus, the regression methods must be chosen appropriately in order for the GHCM to perform well. In Section 4.3 we show that a version of the GHCM incorporating sample splitting has uniform power against alternatives where the expected conditional covariance operator $\mathbb{E}\{\mathrm{Cov}(X, Y \mid Z)\}$ has Hilbert–Schmidt norm of order $n^{-1/2}$, and is thus rate optimal.
4. **The regression methods are only required to perform well on the observed data.** The fact that control of the type I error of the GHCM depends on an in-sample MSPE rather than a more conventional out-of-sample MSPE, has important consequences. While in-sample and out-of-sample errors may be considered rather similar, in the context of function regression, they are substantially different. We demonstrate in Section 4.4 that bounds on the former are achievable under significantly weaker conditions than equivalent bounds on the latter by considering ridge regression in the functional linear model. In particular the required prediction error rates are satisfied over classes of functional linear models where the eigenvalues of the covariance operator of the functional regressor are dominated by a summable sequence; no additional eigen-spacing conditions, or lower bounds on the decay of the eigenvalues are needed, in contrast to existing results on out-of-sample error rates (Cai & Hall, 2006; Crambes & Mas, 2013; Hall & Horowitz, 2007).
5. **The GHCM has several uses.** Section 5 presents the results of numerical experiments on the GHCM. We study the following use cases. (i) Testing for significance of functional predictors

in functional regression models. We are not aware of other approaches that provide significance statements in functional regression models and come with statistical guarantees. For example, in comparison to the p-values from `pfr`, which are highly anti-conservative in challenging setups, the type I error of the GHCM test is well-controlled (see Figure 1). (ii) Deriving confidence intervals for truncation points in truncated functional linear model. We demonstrate in Section 5.2 the use of the GHCM in the construction of a confidence interval for the truncation point in a truncated functional linear model, a problem which we show may be framed as one of testing certain conditional independencies. (iii) Testing for edge presence in functional graphical models. In Section 5.3, we use the GHCM to learn functional graphical models for EEG data from a study on alcoholism.

We conclude with a discussion in Section 6 outlining potential follow-on work and open problems. The supplementary material contains the proofs of all results presented in the main text and some additional numerical experiments, as well as the uniform convergence results mentioned above. An `R`-package `ghcm` (Lundborg et al., 2022) implementing the methodology is available on CRAN.

1.2 | Preliminaries and notation

For three random elements X, Y and Z defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in measurable spaces $(\mathcal{X}, \mathcal{A})$, $(\mathcal{Y}, \mathcal{G})$ and $(\mathcal{Z}, \mathcal{K})$ respectively, we say that X is conditionally independent of Y given Z and write $X \perp\!\!\!\perp Y \mid Z$ when

$$\mathbb{E}(f(X)g(Y) \mid Z) \overset{a.s}{=} \mathbb{E}(f(X) \mid Z)\mathbb{E}(g(Y) \mid Z)$$

for all bounded and Borel measurable $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$. Several equivalent definitions are given in Constantinou and Dawid (2017, Proposition 2.3). As with Euclidean variables, the interpretation of $X \perp\!\!\!\perp Y \mid Z$ is that 'knowing Z renders X irrelevant for predicting Y' (Lauritzen, 1996).

Throughout the paper we consider families of probability distributions $\mathcal{P}$ of the triplet (X, Y, Z), which we partition into the null hypothesis $\mathcal{P}_0$ of those $P \in \mathcal{P}$ satisfying $X \perp\!\!\!\perp Y \mid Z$, and set of alternatives $\mathcal{Q} := \mathcal{P} \setminus \mathcal{P}_0$ where the conditional independence relation is violated. We consider data (x_i, y_i, z_i), $i = 1, \ldots, n$, consisting of i.i.d. copies of (X, Y, Z), and write $X^{(n)} := (x_i)_{i=1}^n$ and similarly for $Y^{(n)}$ and $Z^{(n)}$. We apply to these data a test $\psi_n : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^n \to \{0, 1\}$, with a value of 1 indicating rejection. We will at times write $\mathbb{E}_P(\cdot)$ for expectations of random elements whose distribution is determined by P, and similarly $\mathbb{P}_P(\cdot) = \mathbb{E}_P(\mathbb{1}_{\{\cdot\}})$. Thus, the size of the test ψ_n may be written as $\sup_{P \in \mathcal{P}_0} \mathbb{P}_P(\psi_n = 1)$.

We always take $\mathcal{X} = \mathcal{H}_X$ and $\mathcal{Y} = \mathcal{H}_Y$ for separable Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ and write d_X and d_Y for their dimensions, which may be ∞. When these are finite dimensional, as will typically be the case in practice, $X^{(n)}$ will be a $n \times d_X$ matrix and similarly for $Y^{(n)}$. Similarly, we will take $\mathcal{Z} = \mathbb{R}^{d_Z}$ in the finite-dimensional case and then $Z^{(n)} \in \mathbb{R}^{n \times d_Z}$. However, in order for our theoretical results to be relevant for settings where d_X and d_Y may be arbitrarily large compared to n, our theory must also accommodate infinite-dimensional settings, for which we introduce the following notation.

For g and h in a Hilbert space $\mathcal{H}$, we write $\langle g, h \rangle$ for the inner product of g and h and $\|g\|$ for its norm; note we suppress dependence of the norm and inner product on the Hilbert space. The bounded linear operator on $\mathcal{H}$ given by $x \mapsto \langle x, g\rangle h$ is the outer product of g and h and is denoted by $g \otimes h$. A bounded linear operator $\mathbf{A}$ on $\mathcal{H}$ is compact if it has a singular value decomposition, that is, there exists two orthonormal bases $(e_{1,k})_{k\in\mathbb{N}}$ and $(e_{2,k})_{k\in\mathbb{N}}$ of $\mathcal{H}$ and a non-increasing sequence $(\lambda_k)_{k\in\mathbb{N}}$ of singular values such that

$$\mathbf{A}h = \sum_{k=1}^{\infty} \lambda_k (e_{1,k} \otimes e_{2,k})h = \sum_{k=1}^{\infty} \lambda_k \langle e_{1,k}, h\rangle e_{2,k}$$

for all $h \in \mathcal{H}$. For a compact linear operator $\mathbf{A}$ as above, we denote by $\|\mathbf{A}\|_{\text{op}}$, $\|\mathbf{A}\|_{\text{HS}}$ and $\|\mathbf{A}\|_{\text{TR}}$ the operator norm, Hilbert–Schmidt norm and trace norm, respectively, of $\mathbf{A}$, which equal the ℓ^∞, ℓ^2 and ℓ^1 norms, respectively, of the sequence of singular values $(\lambda_k)_{k\in\mathbb{N}}$.

A random variable on a separable Banach space $\mathcal{B}$ is a mapping $X : \Omega \to \mathcal{B}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which is measurable with respect to the Borel σ-algebra on $\mathcal{B}$, $\mathbb{B}(\mathcal{B})$. Integrals with values in Hilbert or Banach spaces, including expectations, are Bochner integrals throughout. For a random variable X on Hilbert space $\mathcal{H}$, we define the covariance operator of X by

$$\text{Cov}(X) := \mathbb{E}[(X - \mathbb{E}(X)) \otimes (X - \mathbb{E}(X))] = \mathbb{E}(X \otimes X) - \mathbb{E}(X) \otimes \mathbb{E}(X)$$

whenever $\mathbb{E}\|X\|^2 < \infty$. For $h \in \mathcal{H}$ we thus have

$$\text{Cov}(X)h = \mathbb{E}\left(\langle X, h\rangle^2\right) - \mathbb{E}(\langle X, h\rangle)^2.$$

For another random variable Y with $\mathbb{E}\|Y\|^2 < \infty$, we define the cross-covariance operator of X and Y by

$$\text{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}(X)) \otimes (Y - \mathbb{E}(Y))] = \mathbb{E}(X \otimes Y) - \mathbb{E}(X) \otimes \mathbb{E}(Y).$$

We define conditional variants of the covariance operator and cross-covariance operator by replacing expectations with conditional expectations given a σ-algebra or random variable.

2 | THE HARDNESS OF CONDITIONAL INDEPENDENCE TESTING WITH GAUSSIAN FUNCTIONAL DATA

In this section we present a negative result on the possibility of testing for conditional independence with functional data in the idealised setting where all variables are Gaussian. We take $\mathcal{P}$ to consist of distributions of (X, Y, Z) that are jointly Gaussian with injective covariance operator, where X and Z take values in separable Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Z$ respectively with $\mathcal{H}_Z$ infinite-dimensional, and $Y \in \mathbb{R}^{d_Y}$. We note that in the case where $d_Y = 1$ and $\mathcal{H}_X = \mathcal{H}_Z = L^2([0, 1], \mathbb{R})$, each $P \in \mathcal{P}$ admits a representation as a Gaussian scalar-on-function linear model (1) where Y is the scalar response, and functional covariates X, Z and error ε are all jointly Gaussian with $\varepsilon \perp\!\!\!\perp (X, Z)$ (see Proposition 7 in the supplementary material); the settings with $d_Y > 1$ may be thought of equivalently as multi-response versions of this.

For each Q in the set of alternatives $\mathcal{Q}$, we further define $\mathcal{P}_0^Q \subset \mathcal{P}_0$ by

$$\mathcal{P}_0^Q := \{P \in \mathcal{P}_0 : \text{ the marginal distribution of } Z \text{ under } P \text{ and } Q \text{ is the same}\}.$$

Theorem 1 below shows that not only is it fundamentally hard to test the null hypothesis of $\mathcal{P}_0$ against $\mathcal{Q}$ for all dataset sizes n, but restricting to the null $\mathcal{P}_0^Q$ for $Q \in \mathcal{Q}$ presents an equally hard problem.

Theorem 1. *Given alternative $Q \in \mathcal{Q}$ and $n \in \mathbb{N}$, let ψ_n be a test for null hypothesis $\mathcal{P}_0^Q$ against Q. Then we have that the power is at most the size:*

$$\mathbb{P}_Q(\psi_n = 1) \leq \sup_{P \in \mathcal{P}_0^Q} \mathbb{P}_P(\psi_n = 1).$$

An interpretation of this statement in the context of the functional linear model is that regardless of the number of observations n, there is no non-trivial test for the significance of the functional predictor X, even if the marginal distribution of the additional infinite-dimensional predictor Z is known exactly. It is clear that the size of a test over $\mathcal{P}_0$ is at least as large as that over the null $\mathcal{P}_0^Q$, so testing the larger null is of course at least as hard.

It is known that testing conditional independence in simple multivariate (finite-dimensional) settings is hard in the sense of Theorem 1 when the conditioning variable is continuous. In such settings, restricting the null to include only distributions with Lipschitz densities, for example, allows for the existence of tests with power against large classes of the alternative. The functional setting is, however, very different, simply removing pathological distributions from the entire null of conditional independence does not make the problem testable. Even with the parametric restriction of Gaussianity, the null is still too large for the existence of non-trivial hypothesis tests. Indeed, the starting point of our proof is a result due to Kraft (1955) that the hardness in the statement of Theorem 1 is equivalent to the n-fold product $Q^{\otimes n}$ lying in the convex closure in total variation distance of the set of n-fold products of distributions in $\mathcal{P}_0^Q$.

A consequence of Theorem 1 is that we need to make strong modelling assumptions in order to test for conditional independence in the functional data setting. Given the plethora of regression methods for functional data, we argue that it can be convenient to frame these modelling assumptions in terms of regression models for each of X and Y on Z, or more generally, in terms of the performances of methods for these regressions. The remainder of this paper is devoted to developing a family of conditional independence tests whose validity rests primarily on the prediction errors of these regressions.

3 | GHCM METHODOLOGY

In this section we present the Generalised Hilbertian Covariance Measure (GHCM) for testing conditional independence with functional data. To motivate the approach we take, it will be helpful to first review the construction of the Generalised Covariance Measure (GCM) developed in Shah and Peters (2020) for univariate X and Y, which we do in the next section. In Section 3.2 we then define the GHCM.

3.1 Motivation

Consider first therefore the case where X and Y are real-valued random variables, and Z is a random variable with values in some space $\mathcal{Z}$. We can always write $X = f(Z) + \varepsilon$ where $f(z) := \mathbb{E}(X \mid Z = z)$ and similarly $Y = g(Z) + \xi$ with $g(z) := \mathbb{E}(Y \mid Z = z)$. The conditional covariance of X and Y given Z,

$$\operatorname{Cov}(X, Y \mid Z) := \mathbb{E}[\{X - \mathbb{E}(X \mid Z)\}\{Y - \mathbb{E}(Y \mid Z)\} \mid Z] = \mathbb{E}(\varepsilon\xi \mid Z),$$

has the property that $\operatorname{Cov}(X, Y \mid Z) = 0$ and hence $\mathbb{E}(\varepsilon\xi) = 0$ whenever $X \perp\!\!\!\perp Y \mid Z$. The GCM forms an empirical version of $\mathbb{E}(\varepsilon\xi)$ given data $(x_i, y_i, z_i)_{i=1}^n$ by first regressing each of $X^{(n)}$ and $Y^{(n)}$ onto $Z^{(n)}$ to give estimates $\hat{f}$ and $\hat{g}$ of f and g respectively. Using the corresponding residuals $\hat{\varepsilon}_i := x_i - \hat{f}(z_i)$ and $\hat{\xi}_i := y_i - \hat{g}(z_i)$, the product $R_i := \hat{\varepsilon}_i\hat{\xi}_i$ is computed for each $i = 1, \ldots, n$ and then averaged to give $\overline{R} := \sum_{i=1}^n R_i/n$, an estimate of $\mathbb{E}(\varepsilon\xi)$. The standard deviation of $\overline{R}$ under the null $X \perp\!\!\!\perp Y \mid Z$ may also be estimated, and it can be shown (Shah & Peters, 2020, Theorem 8) that under some conditions, $\overline{R}$ divided by its estimated standard deviation converges uniformly to a standard Gaussian distribution.

This basic approach can be extended to the case where X and Y take values in $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$ respectively, by considering a multivariate conditional covariance,

$$\operatorname{Cov}(X, Y \mid Z) := \mathbb{E}\left[\{X - \mathbb{E}(X \mid Z)\}\{Y - \mathbb{E}(Y \mid Z)\}^\top \mid Z\right] = \mathbb{E}(\varepsilon\xi^\top \mid Z) \in \mathbb{R}^{d_X \times d_Y}.$$

This is a zero matrix when $X \perp\!\!\!\perp Y \mid Z$, and hence $\mathbb{E}(\varepsilon\xi^\top) = 0$ under this null. Thus, $\overline{R}$ defined as before but where $R_i := \hat{\varepsilon}_i\hat{\xi}_i^\top$ can form the basis of a test of conditional independence. There are several ways to construct a final test statistic using $\overline{R} \in \mathbb{R}^{d_X \times d_Y}$. The approach taken in Shah and Peters (2020) involves taking the maximum absolute value of a version of $\overline{R}$ with each entry divided by its estimated standard deviation. This, however, does not generalise easily to the functional data setting we are interested in here; we now outline an alternative that can be extended to handle functional data.

To motivate our approach, consider multiplying $\overline{R}$ by $\sqrt{n}$:

$$\begin{aligned}
\sqrt{n}\overline{R} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\varepsilon}_i\hat{\xi}_i^\top = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(z_i) - \hat{f}(z_i) + \varepsilon_i)(g(z_i) - \hat{g}(z_i) + \xi_i)^\top \\
&= \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i\xi_i^\top}_{U_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(z_i) - \hat{f}(z_i))(g(z_i) - \hat{g}(z_i))^\top}_{a_n} \\
&\quad + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(z_i) - \hat{f}(z_i))\xi_i^\top}_{b_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i(g(z_i) - \hat{g}(z_i))^\top}_{c_n}.
\end{aligned} \tag{2}$$

Observe that U_n is a sum of i.i.d. terms and so the multivariate central limit theorem dictates that $U_n/\sqrt{n}$ converges to a $d_X \times d_Y$-dimensional Gaussian distribution. Applying the Frobenius norm $\|\cdot\|_{\mathrm{F}}$ to the a_n term, we get by submultiplicativity and the Cauchy–Schwarz inequality,

$$\|a_n\|_{\mathrm{F}} \leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \|f(z_i) - \hat{f}(z_i)\|_2 \|g(z_i) - \hat{g}(z_i)\|_2$$
$$\leq \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} \|f(z_i) - \hat{f}(z_i)\|_2^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \|g(z_i) - \hat{g}(z_i)\|_2^2 \right)^{1/2}, \tag{3}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. The right-hand side here is a product of in-sample mean squared prediction errors for each of the regressions performed. Under the null of conditional independence, each term of b_n and c_n is mean zero conditional on $(X^{(n)}, Z^{(n)})$ and $(Y^{(n)}, Z^{(n)})$ respectively. Thus, so long as both of the regression functions are estimated at a sufficiently fast rate, we can expect a_n, b_n, c_n to be small so the distribution of $\sqrt{n}\overline{R}$ can be well-approximated by the Gaussian limiting distribution of $U_n/\sqrt{n}$. As in the univariate setting, it is crucially the product of the prediction errors in (3) that is required to be small, so each root mean squared prediction error term can decay at relatively slow $o(n^{-1/4})$ rates.

Unlike the univariate setting, however, $\sqrt{n}\overline{R}$ is now a matrix and hence we need to choose some sensible aggregator function $t : \mathbb{R}^{d_X \times d_Y} \to \mathbb{R}$ such that we can threshold $t(\sqrt{n}\overline{R})$ to yield a p-value. One option is as follows; we take a different approach as the basis of the GHCM for reasons which will become clear in the sequel. If we vectorise $\overline{R}$, that is, view the matrix as a $d_X d_Y$-dimensional vector, then under the assumptions required for the above heuristic arguments to formally hold, $\sqrt{n}\mathrm{Vec}(\overline{R})$ converges to a Gaussian with mean zero and some covariance matrix $C \in \mathbb{R}^{d_X d_Y \times d_X d_Y}$ if $X \perp\!\!\!\perp Y \mid Z$. Provided C is invertible, $\sqrt{n}C^{-1/2}\overline{R}$ therefore converges to a Gaussian with identity covariance under the null and hence $\|C^{-1/2}\sqrt{n}\overline{R}\|_2^2$ converges to a χ^2-distribution with $d_X d_Y$ degrees of freedom. Replacing C with an estimate $\hat{C}$ then yields a test statistic from which we may derive a p-value.

3.2 | The GHCM

We now turn to the setting where X and Y take values in separable Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively. These could for example be $L^2([0,1], \mathbb{R})$, or $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$ respectively, but where X and Y are vectors of function evaluations. The latter case, which we will henceforth refer to as the finite-dimensional case, corresponds to how data would often be received in practice with the observation vectors consisting of function evaluations on fixed grids (which are not necessarily equally spaced). However, it is important to recognise that the dimensions d_X and d_Y of the grids may be arbitrarily large, and it is necessary for the methodology to accommodate this; as we will see, the approach for the multivariate setting described in the previous section does not satisfy this requirement whereas our proposed GHCM will do so.

In some settings, our observed vectors of function evaluations will not be on fixed grids, and the numbers of function evaluations may vary from observation to observation. In Section 3.2.1 we set out a scheme to handle this case and bring it within our framework here.

Similarly to the approach outlined in Section 3.1, we propose to first regress each of $X^{(n)}$ and $Y^{(n)}$ onto $Z^{(n)}$ to give residuals $\hat{\varepsilon}_i \in \mathcal{H}_X$, $\hat{\xi}_i \in \mathcal{H}_Y$ for $i = 1, \ldots, n$. (In practice, these regressions could be performed by `pfr` or `pffr` in the `refund` package (Goldsmith et al., 2011; Ivanescu et al., 2015) or boosting (Brockhaus et al., 2020), for instance.) We centre the residuals, as these and other functional regression methods do not always produce mean-centred residuals. With these residuals we proceed as in the multivariate case outlined above but replacing matrix outer products in the multivariate setting with outer products in the Hilbertian sense, that is we define for $i = 1, \dots, n$,

$$\mathbf{R}_i := \hat{\varepsilon}_i \otimes \hat{\xi}_i, \text{ and } \mathbf{T}_n := \sqrt{n}\,\overline{\mathbf{R}} \quad \text{where } \overline{\mathbf{R}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{R}_i. \tag{4}$$

We can show (see Theorem 2) that under the null, provided the analogous prediction error terms in (3) decay sufficiently fast and additional regularity conditions hold, $\mathbf{T}_n$ above converges uniformly to a Gaussian distribution in the space of Hilbert–Schmidt operators. This comes as a consequence of new results we prove on uniform convergence of Banachian random variables. Moreover, the covariance operator of this limiting Gaussian distribution can be estimated by the empirical covariance operator

$$\hat{\mathbf{C}} := \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{R}_i - \overline{\mathbf{R}}) \otimes_{\mathrm{HS}} (\mathbf{R}_i - \overline{\mathbf{R}}) \tag{5}$$

where $\otimes_{\mathrm{HS}}$ denotes the outer product in the space of Hilbert–Schmidt operators.

An analogous approach to that outlined above for the multivariate setting would involve attempting to whiten this limiting distribution using the square root of the inverse of $\hat{\mathbf{C}}$. However, here we hit a clear obstacle: even in the finite-dimensional setting, whenever $d_X d_Y \geq n$, the inverse of $\hat{\mathbf{C}}$ or $\hat{C}$ from the previous section, cannot exist. Moreover, as indicated by Bai and Saranadasa (1996), who study the problem of testing whether a finite-dimensional Gaussian vector has mean zero, even when the inverses do exist, the estimated inverse covariance may not approximate its population level counterpart sufficiently well. Instead, Bai and Saranadasa (1996) advocate using a test statistic based on the squared ℓ_2-norm of the Gaussian vector.

We take an analogous approach here, and use as our test statistic

$$T_n := \|\mathbf{T}_n\|_{\mathrm{HS}}^2 \tag{6}$$

where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert–Schmidt norm. A further advantage of this test statistic is that it admits an alternative representation given by

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \langle \hat{\varepsilon}_i, \hat{\varepsilon}_j \rangle \langle \hat{\xi}_i, \hat{\xi}_j \rangle; \tag{7}$$

see Section C.1 for a derivation. Only inner products between residuals need to be computed, and so in the finite-dimensional case with the standard inner product, the computational burden is only $O(\max(d_X, d_Y)n^2)$.

As $\mathbf{T}_n$ has an asymptotic Gaussian distribution under the null with an estimable covariance operator, we can deduce the asymptotic null distribution of T_n as a function of $\mathbf{T}_n$. This leads to the α-level test function ψ_n given by

$$\psi_n := \mathbb{1}_{\{T_n \geq q_\alpha\}} \tag{8}$$

where q_α is the $1-\alpha$ quantile of a weighted sum

$$\sum_{k=1}^{d} \lambda_k W_k$$

of independent χ^2_1 distributions $(W_k)_{k=1}^d$ with weights given by the d non-zero eigenvalues $(\lambda_k)_{k=1}^d$ of $\hat{\mathbf{C}}$. Note that $d \leq \min(n-1, d_X d_Y)$.

These eigenvalues may also be derived from inner products of the residuals: they are equal to the eigenvalues of the $n \times n$ matrix

$$\frac{1}{n-1}(\Gamma - J\Gamma - \Gamma J + J\Gamma J)$$

where $J \in \mathbb{R}^{n \times n}$ is a matrix with all entries equal to $1/n$, and $\Gamma \in \mathbb{R}^{n \times n}$ has ijth entry given by

$$\Gamma_{ij} := \langle \hat{\varepsilon}_i, \hat{\varepsilon}_j \rangle \langle \hat{\xi}_i, \hat{\xi}_j \rangle; \tag{9}$$

see Section C.1 of the supplementary material for a derivation. Thus, in the finite-dimensional case, the computation of the eigenvalues requires $O(n^2 \max(d_X, d_Y, n))$ operations. In typical usage therefore, the cost for computing the test statistic given the residuals is dominated by the cost of performing the initial regressions, particularly those corresponding to function-on-function regression. Note that there are several schemes for approximating q_α (Farebrother, 1984; Imhof, 1961; Liu et al., 2009); we use the approach of Imhof (1961) as implemented in the `QuadCompForm` package in `R` (Duchesne & de Micheaux, 2010) in all of our numerical experiments. We summarise the above construction of our test function for the finite-dimensional case with the standard inner product in Algorithm 1.

In principle, different inner products may be chosen, to yield different test functions. However, the theoretical properties of the test function rely on the prediction errors of the regressions, measured in terms of the norm corresponding to the inner product used, being small. In the common case where the observed data are finite vectors of function evaluations, that is, for each $i = 1, \ldots, n$, $x_{ik} = W_{X,i}(k/d_X)$ for a function $W_{X,i} \in L_2([0,1], \mathbb{R})$, and similarly for y_i, our default recommendation is to use the standard inner product. The residuals, $\hat{\varepsilon}_i \in \mathbb{R}^{d_X}$ and $\hat{\xi}_i \in \mathbb{R}^{d_Y}$, would then similarly correspond to underlying functional residuals via $\hat{\varepsilon}_{ik} = W_{\hat{\varepsilon},i}(k/d_X)$ for $W_{\hat{\varepsilon},i} \in L_2([0,1], \mathbb{R})$, and similarly for $\hat{\xi}_i$. We may compare the test function computed based on the computed residuals $\hat{\varepsilon}_i$ and $\hat{\xi}_i$ with that which would be obtained when replacing these with the underlying functions $W_{\hat{\varepsilon},i}$ and $W_{\hat{\xi},i}$. As the test function depends entirely on inner products between residuals, it suffices to compare

$$\hat{\varepsilon}_i^\top \hat{\varepsilon}_j = \sum_{k=1}^{d_X} W_{\hat{\varepsilon},i}(k/d_X) W_{\hat{\varepsilon},i}(k/d_X) \quad \text{and} \quad \int_0^1 W_{\hat{\varepsilon},i}(t) W_{\hat{\varepsilon},j}(t)\, \mathrm{d}t. \tag{10}$$

We see that the LHS is d_X times a Riemann sum approximation to the integral on the RHS. The p-value computed is invariant to multiplicative scaling of the test statistic, and so in the so-called densely observed case where d_X is large, the p-value from the finite-dimensional setting would be a close approximation to that which would be obtained with the true underlying functions.

Other numerical integration schemes could be used to make the approximation even more precise. However, the theory we present in Section 4 that guarantees uniform asymptotic level control and power over certain classes of nulls and alternatives applies directly to the finite-dimensional or infinite-dimensional settings, and so there is no requirement that the approximation error above is small. In particular, there is no strict requirement that the residuals computed correspond to function evaluations on equally spaced grids. However, in that case $\hat{\varepsilon}_i^\top \hat{\varepsilon}_j$ will not necessarily approximate a scaled version of the RHS of (10), and an inner product that maintains this approximation may be more desirable from a power perspective.

In the following section we explain how when the residuals $\hat{\varepsilon}_i$ and $\hat{\xi}_i$ correspond to function evaluations on different grids for each i, we can preprocess these to obtain residuals corresponding to fixed grids, which may then be fed into our algorithm.

An `R`-package `ghcm` (Lundborg et al., 2022) implementing the methodology is available on CRAN.

3.2.1 | Data observed on irregularly spaced grids of varying lengths

We now consider the case where $\hat{\varepsilon}_i \in \mathbb{R}^{d_{X,i}}$ with its kth component given by $\hat{\varepsilon}_{ik} = W_{\hat{\varepsilon},i}(t_{ik})$ for $t_{ik}^X \in [0,1]$, and similarly for $\hat{\xi}_i$. Such residuals would typically be output by regression methods when supplied with functional data $x_i \in \mathbb{R}^{d_{X,i}}$ and $y_i \in \mathbb{R}^{d_{Y,i}}$ corresponding to functional evaluations on grids $(t_{ik})_{k=1}^{d_{X,i}}$ and $(t_{ik})_{k=1}^{d_{Y,i}}$ respectively.

Algorithm 1: Generalised Hilbertian Covariance Measure (GHCM)

```
1  input: $X^{(n)} \in \mathbb{R}^{n \times d_X}$, $Y^{(n)} \in \mathbb{R}^{n \times d_Y}$, $Z^{(n)} \in \mathbb{R}^{n \times d_Z}$ ;
2  options: regression methods for each of the regressions ;
3  begin
4      regress $X^{(n)}$ on $Z^{(n)}$ producing residuals $\hat{\varepsilon}_i \in \mathbb{R}^{d_X}$ for $i = 1, \dots, n$ ;
5      regress $Y^{(n)}$ on $Z^{(n)}$ producing residuals $\hat{\xi}_i \in \mathbb{R}^{d_Y}$ for $i = 1, \dots, n$ ;
6      construct $\Gamma \in \mathbb{R}^{n \times n}$ with entries $\Gamma_{ij} \leftarrow \hat{\varepsilon}_i^\top \hat{\varepsilon}_j \hat{\xi}_i^\top \hat{\xi}_j$ (or more generally via (9)) ;
7      compute test statistic $T_n \leftarrow \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \Gamma_{ij}$ ;
8      set $A \leftarrow \frac{1}{n-1}(\Gamma - J\Gamma - \Gamma J + J\Gamma J)$ where $J \in \mathbb{R}^{n \times n}$ has all entries equal to $1/n$ ;
9      compute the non-zero eigenvalues $\lambda_1, \dots, \lambda_d$ of $A$ (there are at most $n - 1$);
10     compute by numerical integration $p$-value $p \leftarrow \mathbb{P}\left(\sum_{k=1}^d \lambda_k \zeta_k^2 > T_n\right)$, where
         $\zeta_1, \dots, \zeta_d$ are independent standard Gaussian variables ;
11 end
12 output: $p$-value $p$;
```

In order to apply our GHCM methodology, we need to represent these residual vectors by vectors of equal lengths corresponding to fixed grids. Our approach is to construct for each i, natural cubic interpolating splines $\hat{W}_{\hat{\varepsilon},i}$ and $\hat{W}_{\hat{\xi},i}$ corresponding to $\hat{\varepsilon}_i$ and $\hat{\xi}_i$ respectively. We may compute the inner product between these functions in $L_2([0,1],\mathbb{R})$ exactly and efficiently as it is the integral of a piecewise polynomial with the degree in each piece at most 6. This gives us the entries of the matrix Γ (9) which we may then use in lines 7 and following in Algorithm 1. Furthermore, Theorems 3 and 4 apply equally well to the setting considered here provided the residuals are understood as the interpolating splines described above, and the fitted regression functions are defined accordingly as the difference between the observed functional responses these functional residuals.

4 | THEORETICAL PROPERTIES OF THE GHCM

In this section, we provide uniform level control guarantees for the GHCM, and uniform power guarantees for a version incorporating sample splitting; note that we do not recommend the use of the latter in practice but consider it a proxy for the GHCM that is more amenable to theoretical analysis in non-null settings. Before presenting these results, we explain the importance of uniform results in this context, and set out some notation relating to uniform convergence.

4.1 | Background on uniform convergence

In Section 2 we saw that even when $\mathcal{P}$ consists of Gaussian distributions over $\mathcal{H}_X \times \mathbb{R}^{d_Y} \times \mathcal{H}_Z$, we cannot ensure that our test has both the desired size α over $\mathcal{P}_0$ and also non-trivial power properties against alternative distributions in $\mathcal{Q}$. We also have the following related result.

Proposition 1. *Let $\mathcal{H}_Z$ be a separable Hilbert space with orthonormal basis $(e_k)_{k\in\mathbb{N}}$. Let $\mathcal{P}$ be the family of Gaussian distributions for $(X, Y, Z) \in \mathbb{R} \times \mathbb{R} \times \mathcal{H}_Z$ with injective covariance operator and where $(X, Y) \perp\!\!\!\perp (Z_{r+1}, Z_{r+2}, \dots) \mid Z_1, \dots, Z_r$ for some $r \in \mathbb{N}$ and $Z_k := \langle e_k, Z\rangle$ for all $k \in \mathbb{N}$. Let $Q \in \mathcal{Q}$ and recall the definition of $\mathcal{P}_0^Q$ from Section 2. Then, for any test ψ_n,*

$$\mathbb{P}_Q(\psi_n = 1) \le \sup_{P\in\mathcal{P}_0^Q} \mathbb{P}_P(\psi_n = 1).$$

In other words, even if we know a basis $(e_k)_{k\in\mathbb{N}}$ such that in particular the conditional expectations $\mathbb{E}(X \mid Z)$ and $\mathbb{E}(Y \mid Z)$ are sparse in that they depend only on finitely many components $Z_1, \dots, Z_r$ (with $r \in \mathbb{N}$ unknown), and the marginal distribution of Z is known exactly, there is still no non-trivial test of conditional independence.

In this specialised setting, it is however possible to give a test of conditional independence that will, for each *fixed* null hypothesis $P \in \mathcal{P}_0$, yield exact size control and power against all alternatives Q for n sufficiently large. These properties are for example satisfied by the nominal α-level t-test ψ_n^{OLS} for Y in a linear model of X on $Y, Z_1, \dots, Z_{a(n)}$ and an intercept term, for some sequence $a(n) < n-1$ with $a(n) \to \infty$ and $n - a(n) \to \infty$ as $n \to \infty$. Indeed,

$$\sup_{P\in\mathcal{P}_0} \lim_{n\to\infty} \mathbb{P}_P(\psi_n^{\mathrm{OLS}} = 1) = \alpha \quad \text{and} \quad \inf_{Q\in\mathcal{Q}} \lim_{n\to\infty} \mathbb{P}_Q(\psi_n^{\mathrm{OLS}} = 1) = 1; \tag{11}$$

see Section C.2 in the supplementary material for a derivation. This illustrates the difference between pointwise asymptotic level control in the left-hand side of (11), and uniform asymptotic level control given by interchanging the limit and the supremum.

Our analysis instead focuses on proving that the GHCM asymptotically maintains its level uniformly over a subset of the conditional independence null. In order to state our results we first introduce some definitions and notation to do with uniform stochastic convergence. Throughout the remainder of this section we tacitly assume the existence of a measurable space $(\Omega, \mathcal{F})$ whereupon all random quantities are defined. The measurable space is equipped with a family of probability measures $(\mathbb{P}_P)_{P\in\mathcal{P}}$ such that the distribution of (X, Y, Z) under $\mathbb{P}_P$ is P. For a subset $\mathcal{A} \subseteq \mathcal{P}$, we say that a sequence of random variables W_n *converges uniformly in distribution to W over $\mathcal{A}$* and write if

$$W_n \underset{\mathcal{A}}{\overset{\mathcal{D}}{\rightrightarrows}} W \quad \text{if} \quad \lim_{n\to\infty} \sup_{P\in\mathcal{A}} d_{\mathrm{BL}}(W_n, W) = 0,$$

where d_{BL} denotes the bounded Lipschitz metric. We say, W_n *converges uniformly in probability to W over $\mathcal{A}$* and write

$$W_n \underset{\mathcal{A}}{\overset{P}{\rightrightarrows}} W \quad \text{if for any } \epsilon > 0, \quad \lim_{n\to\infty} \sup_{P\in\mathcal{A}} \mathbb{P}_P(\|W_n - W\| \geq \epsilon) = 0.$$

We sometimes omit the subscript $\mathcal{A}$ when it is clear from the context. A full treatment of uniform stochastic convergence in a general setting is given in Section B of the supplementary material. Throughout this section we emphasise the dependence of many of the quantities in Section 3.1 on the distribution of (X, Y, Z) with a subscript P, for example, f_P, ε_P etc.

In Sections 4.2 and 4.3 we present general results on the size and power of the GHCM. We take $\mathcal{P}$ to be the set of all distributions over $\mathcal{H}_X \times \mathcal{H}_Y \times \mathcal{Z}$, and $\mathcal{P}_0$ to be the corresponding conditional independence null. We, however, show properties of the GHCM under smaller sets of distributions $\tilde{\mathcal{P}} \subset \mathcal{P}$ with corresponding null distributions $\tilde{\mathcal{P}}_0 \subset \mathcal{P}_0$, where in particular certain conditions on the quality of the regression procedures on which the test is based are met. In Section 4.1 we consider the special case where the regressions of each of X and Y on Z are given by functional linear models and show that Tikhonov regularised regression can satisfy these conditions. We note that throughout, the dimensions d_X and d_Y may be finite or infinite.

4.2 | Size of the test

In order to state our result on the size of the GHCM, we introduce the following quantities. Let

$$u_P(z) := \mathbb{E}_P(\|\varepsilon_P\|^2 \mid Z = z), \quad v_P(z) := \mathbb{E}_P(\|\xi_P\|^2 \mid Z = z).$$

We further define the in-sample unweighted and weighted mean squared prediction errors of the regressions as follows:

$$M^f_{n,P} := \frac{1}{n}\sum_{i=1}^{n} \|f_P(z_i) - \hat{f}^{(n)}(z_i)\|^2, \quad M^g_{n,P} := \frac{1}{n}\sum_{i=1}^{n} \|g_P(z_i) - \hat{g}^{(n)}(z_i)\|^2, \tag{12}$$

$$\tilde{M}^f_{n,P} := \frac{1}{n}\sum_{i=1}^{n} \|f_P(z_i) - \hat{f}^{(n)}(z_i)\|^2 v_P(z_i), \quad \tilde{M}^g_{n,P} := \frac{1}{n}\sum_{i=1}^{n} \|g_P(z_i) - \hat{g}^{(n)}(z_i)\|^2 u_P(z_i). \tag{13}$$

The result below shows that on a subset $\tilde{\mathcal{P}}_0$ of the null distinguished primarily by the product of the prediction errors in (12) being small, the operator-valued statistic $\mathbf{T}_n$ converges in distribution uniformly to a mean zero Gaussian whose covariance can be estimated consistently. We remark that prediction error quantities in (12) and (13) are ‘in-sample’ prediction errors, only reflecting the quality of estimates of the conditional expectations f and g at the observed values $z_1, \ldots, z_n$.

Theorem 2. *Let $\tilde{\mathcal{P}}_0 \subseteq \mathcal{P}_0$ be such that uniformly over $\tilde{\mathcal{P}}_0$,*

1. $nM^f_{n,P}M^g_{n,P} \overset{P}{\rightrightarrows} 0$,
2. $\tilde{M}^f_{n,P} \overset{P}{\rightrightarrows} 0, \tilde{M}^g_{n,P} \overset{P}{\rightrightarrows} 0$,
3. $\inf_{P\in\tilde{\mathcal{P}}_0} \mathbb{E}_P(\|\varepsilon_P\|^2\|\xi_P\|^2) > 0$ *and* $\sup_{P\in\tilde{\mathcal{P}}_0} \mathbb{E}_P(\|\varepsilon_P\|^{2+\eta}\|\xi_P\|^{2+\eta}) < \infty$ *for some* $\eta > 0$, *and*
4. *for some orthonormal bases* $(e_{X,i})_{i=1}^{d_X}$ *and* $(e_{Y,j})_{j=1}^{d_Y}$ *of* $\mathcal{H}_X$ *and* $\mathcal{H}_Y$, *respectively, writing* $\varepsilon_{P,i} := \langle e_{X,i}, \varepsilon_P\rangle$ and $\xi_{P,j} := \langle e_{Y,j}, \xi_P\rangle$, *we have*

$$\lim_{K\to\infty} \sup_{P\in\tilde{\mathcal{P}}_0} \sum_{(i,j):i+j\geq K} \mathbb{E}_P(\varepsilon^2_{P,i}\xi^2_{P,j}) = 0,$$

where we interpret an empty sum as 0.

Then uniformly over $\tilde{\mathcal{P}}_0$ we have

$$\mathbf{T}_n \overset{\mathcal{D}}{\rightrightarrows} \mathcal{N}(0, \mathbf{C}_P) \quad \textit{and} \quad \|\hat{\mathbf{C}} - \mathbf{C}_P\|_{\mathrm{TR}} \overset{P}{\rightrightarrows} 0,$$

where

$$\mathbf{C}_P := \mathbb{E}\{(\varepsilon_P \otimes \xi_P) \otimes_{\mathrm{HS}} (\varepsilon_P \otimes \xi_P)\}.$$

Condition (i) is the most important requirement, and says that the regression methods must perform sufficiently well, uniformly on $\tilde{\mathcal{P}}_0$. It is satisfied if $\sqrt{n}M^f_{n,P}$, $\sqrt{n}M^g_{n,P} \overset{P}{\rightrightarrows} 0$, and so allows for relatively slow $o(\sqrt{n})$ rates for the mean squared prediction errors. Moreover, if one regression yields a faster rate, the other can go to zero more slowly. These properties are shared with the regular generalised covariance measure and more generally doubly robust procedures popular in the literature on causal inference and semiparametric statistics (Chernozhukov et al., 2018; Robins & Rotnitzky, 1995; Scharfstein et al., 1999). Condition (ii) is much milder, and if the conditional variances u_P and v_P are bounded almost surely, it is satisfied when simply $M^f_{n,P}$, $M^g_{n,P} \overset{P}{\rightrightarrows} 0$. We note that importantly, the regression methods are not required to extrapolate well beyond the observed data. We show in Section 4.4 that when the regression models are functional linear models and ridge regression is used for the functional regressions, (i) and (ii) hold under much weaker conditions than are typically required for out-of-sample prediction error guarantees in the literature.

Conditions (iii) and (iv) imply that the family $\{\varepsilon_P \otimes \xi_P : P \in \tilde{\mathcal{P}}_0\}$ is uniformly tight. Similar tightness conditions are required in Chen and White (1998, lemma 3.1) in the context of functional central limit theorems. Note that if d_X and d_Y are both finite, this condition is always satisfied.

The result below shows that the GHCM test ψ_n (8) has type I error control uniformly over $\tilde{\mathcal{P}}_0$ given in Theorem 2, provided an additional assumption of non-degeneracy of the covariance operators is satisfied.

Theorem 3. *Let $\tilde{\mathcal{P}}_0 \subseteq \mathcal{P}_0$ satisfy the conditions stated in Theorem 2, and in addition suppose*

$$\inf_{P\in\tilde{\mathcal{P}}_0} \|\mathbf{C}_P\|_{\mathrm{op}} > 0. \tag{14}$$

Then for each $\alpha \in (0,1)$, the α-level GHCM test ψ_n (8) satisfies

$$\lim_{n\to\infty}\sup_{P\in\tilde{\mathcal{P}}_0}|\mathbb{P}_P(\psi_n = 1) - \alpha| = 0. \tag{15}$$

4.3 | Power of the test

We now study the power of the GHCM. It is not straightforward to analyse what happens to the test statistic T_n when the null hypothesis is false in the setup we have considered so far. However, if we modify the test such that the regression function estimates $\hat{f}$ and $\hat{g}$ are constructed using an auxiliary dataset independent of the main data $(x_i, y_i, z_i)_{i=1}^n$, the behaviour of T_n is more tractable. Given a single sample, this could be achieved through sample splitting, and cross-fitting (Chernozhukov et al., 2018) could be used to recover the loss in efficiency from the split into smaller datasets. However, we do not recommend such sample splitting in practice here and view this as more of a technical device that facilitates our theoretical analysis. As we require $\hat{f}$ and $\hat{g}$ to satisfy (i) and (ii) of Theorem 2, these estimators would need to perform well out of sample rather than just on the observed data, which is typically a harder task.

Given that our test is based on an empirical version of $\mathbb{E}(\mathrm{Cov}(X, Y \mid Z)) = \mathbb{E}(\varepsilon \otimes \xi)$, we can only hope to have power against alternatives where this is non-zero. For such alternatives, however, we have positive power whenever the Hilbert–Schmidt norm of the expected conditional covariance operator is at least $c/\sqrt{n}$ for a constant $c > 0$, as the following result shows.

Theorem 4. *Consider a version of the GHCM test ψ_n where $\hat{f}$ and $\hat{g}$ are constructed on independent auxiliary data. Let $\tilde{\mathcal{P}} \subset \mathcal{P}$ be the set of distributions for (X, Y, Z) satisfying (i)–(iv) of Theorem 2 and (14) with $\tilde{\mathcal{P}}$ in place of $\tilde{\mathcal{P}}_0$. Then writing $\mathbf{K}_P := \mathbb{E}_P(\varepsilon_P \otimes \xi_P) = \mathbb{E}_P(\mathrm{Cov}_P(X, Y \mid Z))$, we have, uniformly over $\tilde{\mathcal{P}}$,*

$$\tilde{\mathbf{T}}_n := \frac{1}{\sqrt{n}}\sum_{i=1}^n(\mathbf{R}_i - \mathbf{K}_P) \overset{\mathcal{D}}{\rightrightarrows} \mathcal{N}(0, \mathbf{C}_P) \quad \text{and} \quad \|\hat{\mathbf{C}} - \mathbf{C}_P\|_{\mathrm{TR}} \overset{P}{\rightrightarrows} 0.$$

Furthermore, an α-level GHCM test ψ_n (constructed using independent estimates $\hat{f}$ and $\hat{g}$) satisfies the following two statements.

1. *Redefining $\tilde{\mathcal{P}}_0 = \tilde{\mathcal{P}} \cap \mathcal{P}_0$, we have that (15) is satisfied, and so an α-level GHCM test has size converging to α uniformly over $\tilde{\mathcal{P}}_0$.*
2. *For every $0 < \alpha < \beta < 1$ there exists $c > 0$ and $N \in \mathbb{N}$ such that for any $n \geq N$,*

$$\inf_{P\in Q_{c,n}} \mathbb{P}_P(\psi_n = 1) \geq \beta,$$

where $Q_{c,n} := \{P \in \tilde{\mathcal{P}} \ : \ \|\mathbf{K}_P\|_{\mathrm{HS}} > c/\sqrt{n}\}$.

In a setting where X, Y and Z are related by linear regression models, we can write down $\|\mathbb{E}\mathrm{Cov}(X, Y \mid Z)\|_{\mathrm{HS}}$ more explicitly. Suppose Z, ε and ξ are independent random variables in $L^2([0,1], \mathbb{R})$, with X and Y determined by

$$X(t) = \int \beta^X(s,t)Z(s)\,\mathrm{d}s + \varepsilon(t)$$

$$Y(t) = \int \beta^Y(s,t)Z(s)\,\mathrm{d}s + \int \theta(s,t)X(s)\,\mathrm{d}s + \varepsilon + \xi(t).$$

Then $\mathbb{E}\mathrm{Cov}(X, Y \mid Z)$ is an integral operator with kernel

$$\phi(s,t) = \int_0^1 \theta(u,s)\nu(t,u)\,\mathrm{d}u,$$

where $\nu(t,u)$ denotes the covariance function of ε. The Hilbert–Schmidt norm $\|\mathbb{E}\mathrm{Cov}(X, Y \mid Z)\|_{\mathrm{HS}}$ is then given by the $L^2([0,1]^2, \mathbb{R})$-norm of ϕ. We investigate the empirical performance of the GHCM in such a setting in Section 5.1.2.

4.4 | GHCM using linear function-on-function ridge regression

Here we consider a special case of the general setup used in Sections 4.2 and 4.3 where we assume that $\mathcal{Z}$ is a Hilbert space $\mathcal{H}_Z$ and that, under the null of conditional independence, the Hilbertian X and Y are related to Hilbertian Z via linear models:

$$X = \mathbf{S}_P^X Z + \varepsilon_P \tag{16}$$

$$Y = \mathbf{S}_P^Y Z + \xi_P. \tag{17}$$

Here $\mathbf{S}_P^X$ is a Hilbert–Schmidt operator such that $\mathbf{S}_P^X Z = f(Z) := \mathbb{E}(X \mid Z)$, with analogous properties holding for $\mathbf{S}_P^Y$, and it is assumed that $\mathbb{E}Z = 0$. If X, Y and Z are elements of $L^2([0,1], \mathbb{R})$, this is equivalent to

$$X(t) = \int_0^1 \beta_P^X(s,t)Z(s)\,\mathrm{d}s + \varepsilon_P(t), \tag{18}$$

where β_P^X is a square-integrable function, and similarly for the relationship between Y and Z. Such functional response linear models have been discussed by Ramsay and Silverman (2005, chapter 16), and studied by Chiou et al. (2004), Yao et al. (2005), Crambes and Mas (2013), for example. Benatia et al. (2017) propose a Tikhonov regularised estimator analogous to ridge regression (Hoerl & Kennard, 2000); applied to the regression model (16), this estimator takes the form

$$\hat{\mathbf{S}} = \underset{\mathbf{S}}{\mathrm{argmin}} \left(\sum_{i=1}^n \|x_i - \mathbf{S}(z_i)\|^2 + \gamma \|\mathbf{S}\|_{\mathrm{HS}}^2 \right), \tag{19}$$

where $\gamma > 0$ is a tuning parameter.

We now consider a specific instance of the general GHCM framework using regression estimates based on (19). Specifically, we form estimate $\hat{\mathbf{S}}^X$ of $\mathbf{S}^X$ by solving the optimisation in (19) with regularisation parameter

$$\hat{\gamma} := \underset{\gamma > 0}{\mathrm{argmin}} \left(\frac{1}{\gamma n} \sum_{i=1}^n \min(\hat{\mu}_i/4, \gamma) + \frac{\gamma}{4} \right), \tag{20}$$

where $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \cdots \geq \hat{\mu}_n \geq 0$ are the ordered eigenvalues of the $n \times n$ matrix K with $K_{ij} = \langle z_i, z_j \rangle / n$. We form estimate $\hat{\mathbf{S}}^Y$ of $\mathbf{S}^Y$ analogously but with the x_i replaced by y_i in (19). Note that in the case where $K = 0$ and so $\hat{\gamma}$ does not exist, we simply take $\hat{\mathbf{S}}^X$ and $\hat{\mathbf{S}}^Y$ to be 0 operators, that is, no regression is performed.

The data-driven choice of $\hat{\gamma}$ above is motivated by an upper bound on the in-sample MSPE of the estimators $\hat{\mathbf{S}}^X$ and $\hat{\mathbf{S}}^Y$ (see Lemma 17 in the supplementary material) where we have omitted some distribution-dependent factors of $\|\mathbf{S}_P^X\|_{\mathrm{HS}}^2$ or $\|\mathbf{S}_P^Y\|_{\mathrm{HS}}^2$ and a variance factor; a similar strategy was used in an analysis of kernel ridge regression (Shah & Peters, 2020) which closely parallels ours here. This choice allows us to conduct a theoretical analysis that we present below. In practice, other choices of regularisation parameter such as cross validation-based approaches may perform even better and so could alternative methods that are not based on Tikhonov regularisation.

In the following result, we take ψ_n to be the α-level GHCM test (8) with estimated regression functions $\hat{f}$ and $\hat{g}$ yielding fitted values given by

$$\hat{f}(z_i) = \hat{\mathbf{S}}^X z_i \quad \text{and} \quad \hat{g}(z_i) = \hat{\mathbf{S}}^Y z_i, \quad \text{for all } i = 1, \cdots, n. \tag{21}$$

Note that in the finite dimensional setting where $X^{(n)} \in \mathbb{R}^{n \times d_X}$ (which is also covered by the result below), we have that the matrix of fitted values $(\hat{f}(z_i))_{i=1}^n \in \mathbb{R}^{n \times d_X}$ is given by

$$K(K + \gamma I)^{-1} X^{(n)},$$

and similarly for the $Y^{(n)}$ regression.

Theorem 5. *Let* $\tilde{\mathcal{P}}_0 \subset \mathcal{P}_0$ *be such that* (16) *and* (17) *are satisfied and moreover* (iii) *and* (iv) *of Theorem 2 and* (14) *hold when* $\hat{f}$ *and* $\hat{g}$ *are as in* (21). *Suppose further that*

1. $\sup_{P \in \tilde{\mathcal{P}}_0} \max(\|\mathbf{S}_P^X\|_{\mathrm{HS}}, \|\mathbf{S}_P^Y\|_{\mathrm{HS}}) < \infty$,
2. $\sup_{P \in \tilde{\mathcal{P}}_0} \max(u_P(Z), v_P(Z)) < \infty$ *almost surely*,
3. $\sup_{P \in \tilde{\mathcal{P}}_0} \mathbb{E}\|Z\|^2 < \infty$ *and* $\lim_{\gamma \downarrow 0} \sup_{P \in \tilde{\mathcal{P}}_0} \sum_{k=1}^{\infty} \min(\mu_{k,P}, \gamma) = 0$ *where* $(\mu_{k,P})_{k \in \mathbb{N}}$ *denote the ordered eigenvalues of the covariance operator of* Z *under* P.

Then the α*-level GHCM test* ψ_n *satisfies*

$$\lim_{n \to \infty} \sup_{P \in \tilde{\mathcal{P}}_0} |\mathbb{P}_P(\psi_n = 1) - \alpha| = 0.$$

Condition (iii) is generally satisfied, by the dominated convergence theorem, for any family $\tilde{\mathcal{P}}_0$ for which the sequence of eigenvalues of the covariance operators are uniformly bounded above by a summable sequence. As a very simple example where all the remaining conditions of Theorem 5 are satisfied, we may consider the family of distribution $\tilde{\mathcal{P}}_0$ where Z, ε_P in (22) and ξ_P in (23) are independent, and the latter two are Brownian motions with variances $\sigma_{\varepsilon,P}^2$ and $\sigma_{\xi,P}^2$ respectively. If the coefficient functions β_P^X corresponding to X in (18) are in $L_2([0,1]^2, \mathbb{R})$ with norms bounded above for all $P \in \mathcal{P}_0$, and an equivalent assumption for the coefficient functions relating to Y holds, and $\sigma_{\varepsilon,P}^2$ and $\sigma_{\xi,P}^2$ are bounded from above and below uniformly, we have that $\mathcal{P}_0$ satisfies all the requirements of Theorem 5.

The proof of Theorem 5 relies on Lemma 17 in Section C.5 of the supplementary material, which gives a bound on the in-sample MSPE of ridge regression in terms of the decay of the eigenvalues $\mu_{k,P}$, which may be of independent interest. For example, we have that if these are dominated by an exponentially decaying sequence, the in-sample MSPE is $o(\log n/n)$ as $n \to \infty$ (see Corollary 2). This matches the out-of-sample MSPE bound obtained in Crambes and Mas (2013), corollary 5 in the same setting as that described, but the out-of-sample result additionally requires convexity and lower bounds on the decay of the sequence of eigenvalues of the covariance operator, and stronger moment assumptions on the norm of the predictor. Similarly, other related results (e.g. Cai & Hall, 2006; Hall & Horowitz, 2007) require additional eigen-spacing conditions in place of convexity, and upper and lower bounds on the decay of the eigenvalues. Furthermore, while some of these bounds are uniform over values of the linear coefficient operator for fixed distributions of the predictors, our in-sample MSPE bound is uniform over both the coefficients and distributions of the predictor. This illustrates how in-sample and out-of-sample prediction are very different in the functional data setting, and reliance on the former being small, as we have with the GHCM, is desirable due to the weaker conditions needed to guarantee this.

5 | EXPERIMENTS

In this section we present the results of numerical experiments that investigate the performance of our proposed GHCM methodology. We implement the GHCM as described in Algorithm 1 with scalar-on-function and function-on-function regressions performed using the `pfr` and `pffr` functions respectively from the `refund` package Goldsmith et al. (2020). These are functional linear regression methods which rely on fitting smoothers implemented in the `mgcv` package (Wood, 2017); we choose the tuning parameters for these smoothers (dimension of the basis expansions of the smooth terms) as per the standard guidance such that a further increase does not decrease the deviance. In Section 5.3 in the supplement, we study high-dimensional EEG data using the GHCM with regressions performed using `FDboost`.

We note that, to the best of our knowledge, neither `FDboost` nor the regression methods in `refund` come with prediction error bounds (such as the ones derived in Section 4.4) that are required for obtaining formal guarantees for the GHCM; nevertheless they are well-developed and well-used functional regression methods and our aim here is to demonstrate empirically that they perform suitably well in terms of prediction such that when used with the GHCM, type I error is maintained across a variety of settings. In Section D of the supplementary material, we include additional simulations that consider among others, settings with heavy tailed errors, test the GHCM with `FDboost` in further settings and examine the local power of the GHCM.

5.1 | Size and power simulation

In this section we examine the size and power properties of the GHCM when testing the conditional independence $X \perp\!\!\!\perp Y \mid Z$. We take $X, Z \in L^2([0,1], \mathbb{R})$, and first consider the setting where Y is scalar. In Section 5.1.2 we present experiments for the case where $Y \in L^2([0,1], \mathbb{R})$, so all variables are functional. All simulated functional random variables are sampled on an equidistant grid of [0,1] with 100 grid points.

5.1.1 | Scalar Y, functional X and Z

Here we consider the setup where Z is standard Brownian motion and X and Y are related to Z through the functional linear models

$$X(t) = \int_0^1 \beta_a(s,t)Z(s)\,\mathrm{d}s + N_X(t), \tag{22}$$

$$Y = \int_0^1 \alpha_a(t)Z(t)\,\mathrm{d}t + N_Y. \tag{23}$$

The variables N_X, N_Y and Z are independent with N_X a Brownian motion with variance σ_X^2, $N_Y \sim \mathcal{N}(0,1)$, so $X \perp\!\!\!\perp Y \mid Z$. Nonlinear coefficient functions β_a and α_a are given by

$$\beta_a(s,t) = a\exp(-(st)^2/2)\sin(ast), \qquad \alpha_a(t) = \int_0^1 \beta_a(s,t)\,\mathrm{d}s. \tag{24}$$

We vary the parameters $\sigma_X \in \{0.1, 0.25, 0.5, 1\}$ and $a \in \{2, 6, 12\}$. We generate n i.i.d. observations from each of the $4 \times 3 = 12$ models given by (22), (23), for sample sizes $n \in \{100, 250, 500, 1000\}$. Increasing a or decreasing σ_X increase the difficulty of the testing problem: for large a, β_a oscillates more, making it harder to remove the dependence of X on Z. A smaller σ_X makes Y closer to the integral of X, and so increases the marginal dependence of X and Y.

We apply the GHCM and compare the resulting tests to those corresponding to the significance test for X in a regression of Y on (X, Z) implemented in `pfr`. The rejection rates of the two tests at the 5% level, averaged over 100 simulation runs, can be seen in Figure 1. We see that the `pfr`

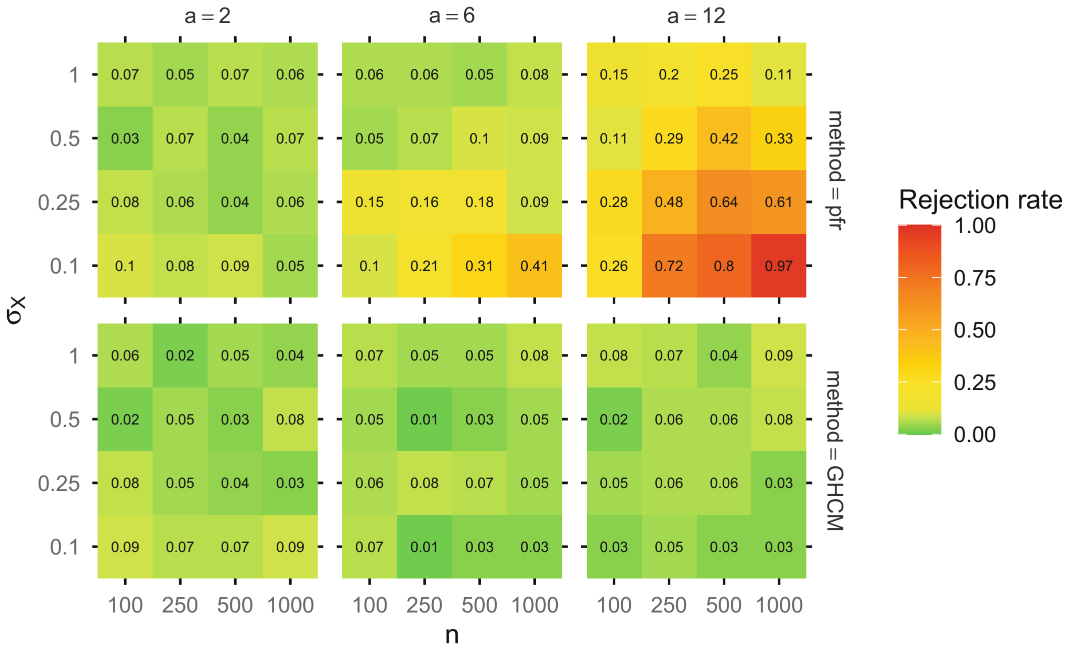

FIGURE 1 Rejection rates in the various null settings considered in Section 5.1.1 for the nominal 5%-level `pfr` test (top) and GHCM test (bottom). [Colour figure can be viewed at wileyonlinelibrary.com]

test has size greatly exceeding its level in the more challenging large a, small σ_X settings, with large values of n exposing most clearly the miscalibration of the test statistic. In these settings, Y may be approximated simply by the integral of X reasonably well, and is also well-approximated by the true regression function that features only Z. Regularisation encourages `pfr` to fit a model where X determines the response, rather than X, and the p-values reflect this. On the other hand, the GHCM tests maintain reasonable type I error control across the settings considered here.

To investigate the power properties of the test, we simulate Z as before with X also generated according to (22). We replace the regression model (23) for Y with

$$Y = \int_0^1 \alpha_a(t) Z(t)\,\mathrm{d}t + \int_0^1 \frac{\alpha_a(t)}{a} X(t)\,\mathrm{d}t + N_Y, \tag{25}$$

where $N_Y \sim \mathcal{N}(0, 1)$ as before. Note that the coefficient function for X oscillates more as a increases. The rejection rates at the 5% level can be seen in Figure 2. While the two approaches perform similarly when $a = 2$, the `pfr` test has higher power in the more complex cases. However, as the results from the size analysis in Figure 1 show, null cases are also rejected in the analogous settings.

To illustrate the full distribution of p-values from the two methods under the null and the alternative, we plot false positive rates and true positive rates in each setting as a function of the chosen significance level of the test α. The full set of results can be seen in Section D of the supplementary material and a plot for a subset of the simulations settings where $n = 500$ and $\sigma_X \in \{0.1, 0.25, 0.5\}$ is presented in Figure 3. We see that both tests distinguish null from alternative well in the cases with a small and σ_X large. The p-values of the GHCM are close to uniform in the settings considered, whereas the distribution of the `pfr` p-values is heavily dependent on the particular null setting, illustrating the difficulty with calibrating this test.

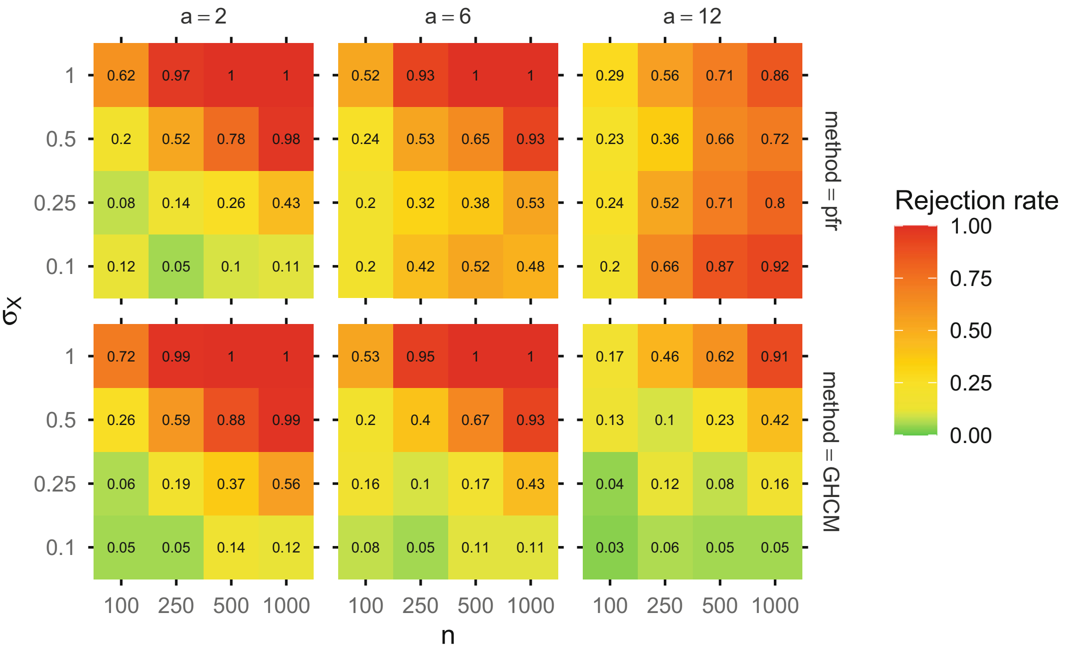

FIGURE 2 Rejection rates in the various alternative settings considered in Section 5.1.1 (see (25)) for the nominal 5%-level `pfr` test (top) and GHCM test (bottom). [Colour figure can be viewed at wileyonlinelibrary.com]

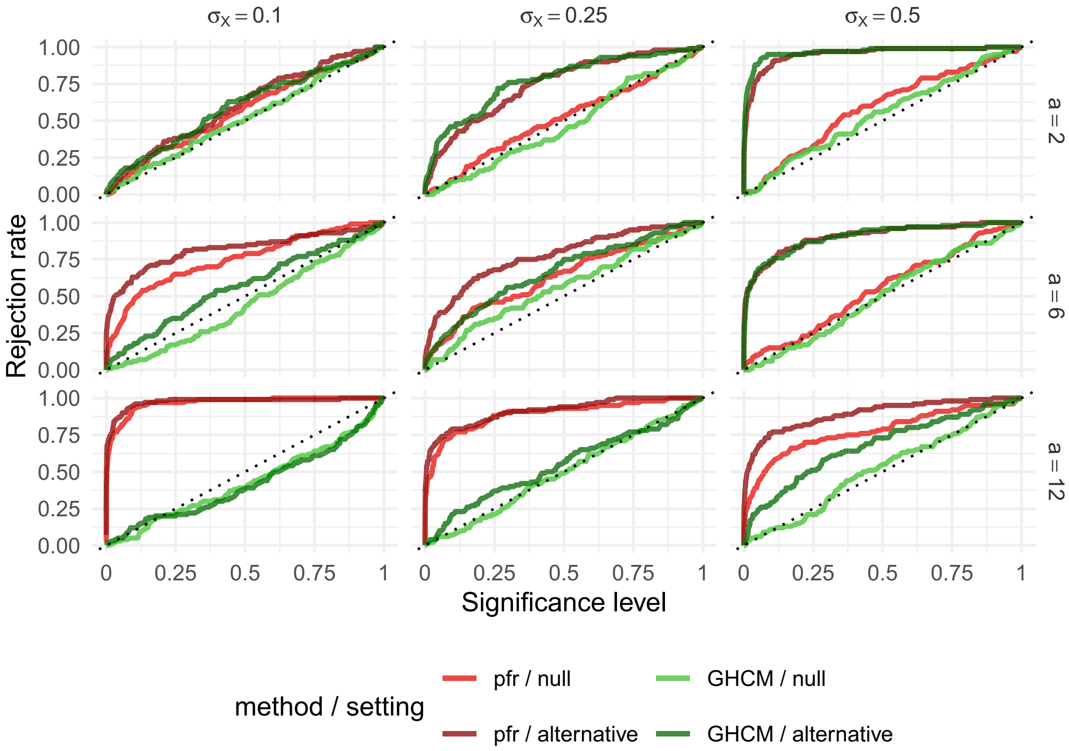

FIGURE 3 Rejection rates against significance level for the `pfr` (red) and GHCM (green) tests under null (light) and alternative (dark) settings when $n = 500$. [Colour figure can be viewed at wileyonlinelibrary.com]

In Section D of the supplementary material we also present the results of two additional sets of experiments. We repeat the experiments above using the `FDboost` package for regressions in place of the `refund` package. We see that the performance of the GHCM with `FDboost` is broadly similar to that displayed in Figures 1 and 2, supporting our theoretical results which indicate that provided the prediction errors of the regression methods used are sufficiently small, the test will perform similarly.

We also consider the case where the noise is heavy tailed. Specifically, we present analogous plots for setting where N_Y is t-distributed with different degrees of freedom, $n = 500$ and $\sigma_X = 0.25$; the results are similar to Figure 3, with the GHCM maintaining type I error control, and `pfr` tending to be anti-conservative in the more challenging settings.

5.1.2 | Functional X, Y and Z

In this section we modify the setup and consider functional $Y \in L^2([0,1],\mathbb{R})$. We take X and Z as in Section 5.1.1 but in the null settings we let

$$Y(t) = \int_0^1 \beta_a(s,t)Z(s)\,\mathrm{d}s + N_Y(t),$$

where N_Y is a standard Brownian motion. Note that this is a particularly challenging setting to maintain type I error control as X and Y are then highly correlated, and moreover the biases from

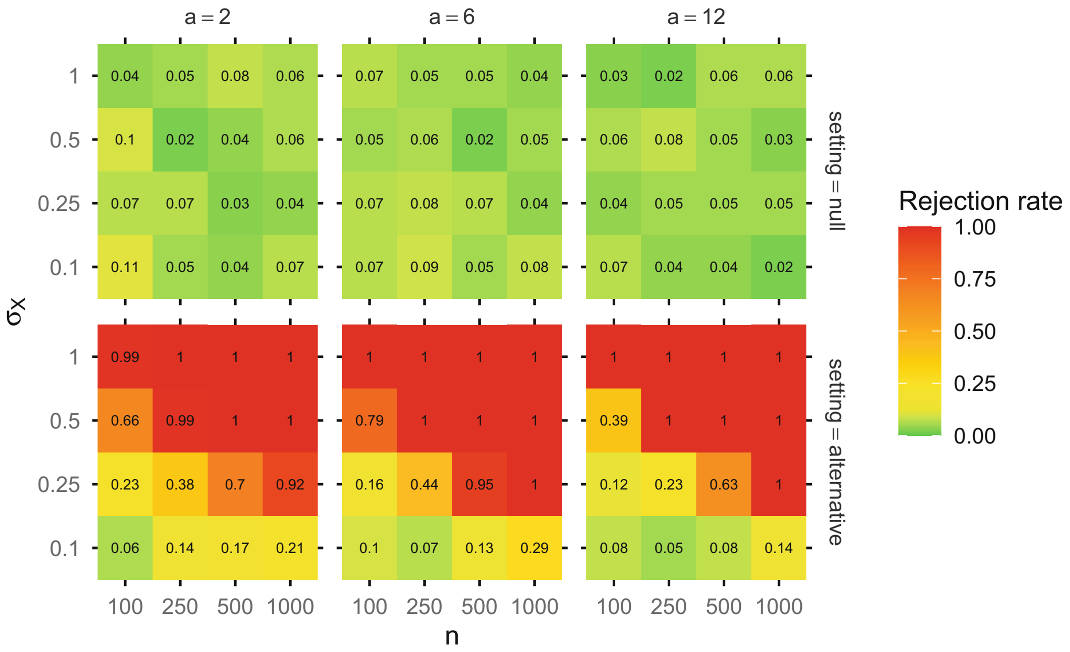

FIGURE 4 Rejection rates in the various null (top) and alternative (bottom) settings considered in Section 5.1.2 for the nominal 5%-level GHCM test. [Colour figure can be viewed at wileyonlinelibrary.com]

regressing each of X and Y on Z will tend to be in similar directions making the equivalent of the term a_n in (2) potentially large.

In the alternative settings, we take

$$Y(t) = \int_0^1 \beta_a(s,t)Z(s)\,\mathrm{d}s + \int_0^1 \frac{\beta_a(s,t)}{a}X(s)\,\mathrm{d}s + N_Y(t)$$

with N_Y again being a standard Brownian motion.

The rejection rates at the 5% level, averaged over 100 simulation runs, can be seen in Figure 4. We see that, as in the case where $Y \in \mathbb{R}$, the GHCM maintains good type I error control in the settings considered, and has power increasing with n and σ_X as expected. We note that a comparison with the p-values from `ff`-terms in the `pffr`-function of the `refund` package here does not seem helpful. In our experiments the corresponding tests consistently reject in true null settings even for simple models.

In Section D of the supplementary material we look at the subset of the settings considered above with $n = 500$ and $\sigma_X = 0.25$ but where X and Y are observed on irregular grids of varying length grids. We first preprocess the residuals output by the regression method as described in Section 3.2.1 and then apply the GHCM. We observe that the performance is similar to that in the fixed grid setting, although the power is lower when the average grid length is smaller, and type I error increases slightly above nominal levels in the most challenging $a = 12$ setting.

5.2 | Confidence intervals for truncated linear models

In this section we consider an application of the GHCM in constructing a confidence interval for the truncation point $\theta \in [0, 1]$ in a truncated functional linear model (Hall & Hooker, 2016)

$$Y = \int_0^\theta \alpha(t)X(t)\,\mathrm{d}t + \varepsilon, \tag{26}$$

where the predictor $X \in L^2([0,1],\mathbb{R})$, $Y \in \mathbb{R}$ is a response and $\varepsilon \perp\!\!\!\perp X$ is stochastic noise. To frame this as a conditional independence testing problem, observe that (26) implies that defining the null hypotheses

$$H_{\tilde\theta} \,:\, Y \perp\!\!\!\perp \{X(t)\}_{t>\tilde\theta} \mid \{X(t)\}_{t\le\tilde\theta} \tag{27}$$

for $\tilde\theta \in (0,1)$, we have that $H_{\tilde\theta}$ is true for all $\theta \le \tilde\theta \le 1$.

Given an α-level conditional independence test ψ, we may thus form a one-sided confidence interval for θ using

$$\left[\inf\{\tilde\theta \in (0,1) \,:\, \psi \text{ accepts null } H_{\tilde\theta}\},\ 1\right]. \tag{28}$$

Indeed, with probability $1-\alpha$, ψ will not reject the true null H_θ, and so with probability $1-\alpha$ the infimum above will be at most θ.

To approximate (28) we initially consider the null hypothesis $H_{\tilde\theta}$ at five equidistant values of $\tilde\theta$ and then employ a bisection search between the smallest of these points $\tilde\theta$ at which $H_{\tilde\theta}$ is accepted by a 5% level GHCM, and the point immediately before it or 0. We consider two instances of the model (26) with $\theta = 0.275, 0.675$ and with $\alpha(t) := 10(t+1)^{-1/3}$, X a standard Brownian motion and $\varepsilon \sim \mathcal{N}(0,1)$. The simulated functional variables are observed on an equidistant grid of $[0,1]$ with 121 grid points. The results across 500 simulations are given in Figure 5. We see that the empirical coverage probabilities are close to the nominal coverage of 95%.

5.3 | EEG data analysis

In this section we demonstrate the application of our GHCM methodology to the problem of learning functional graphical models. In contrast to existing work (Qiao et al., 2019, 2020) which

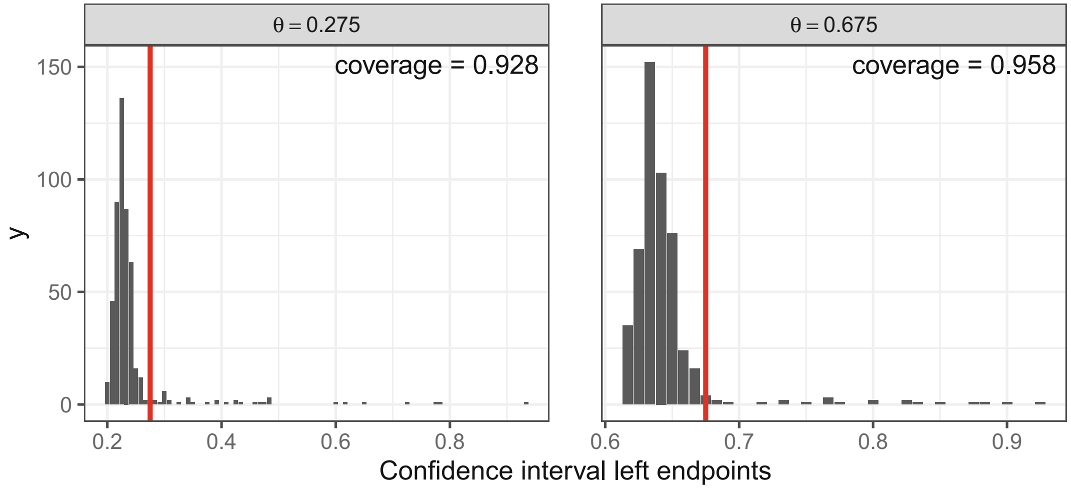

FIGURE 5 Histograms of the left endpoints of 95% confidence intervals for truncation points $\theta = 0.275$ (left) and $\theta = 0.675$ (right), given by red vertical lines, in model (26) across 500 simulations. [Colour figure can be viewed at wileyonlinelibrary.com]

typically assumes a Gaussian functional graphical model and outputs a point estimate of the conditional independence graph, here we are able to test for the presence of each edge, with type I error control guaranteed for data generating processes where our regression methods perform suitably well as indicated by Theorem 3.

We illustrate this on an EEG dataset from a study on alcoholism (Zhang et al., 1995; Ingber, 1997, 1998). The study participants were shown one of three visual stimuli repeatedly and simultaneous EEG activity was measured across 64 channels over the course of 1 second at 256 measurements per second. While the study included both a control group and an alcoholic group we will restrict our analysis to the alcoholic group consisting of 77 subjects and further restrict ourselves to a single type of visual stimulus. We preprocess the data as in Qiao et al. (2019), averaging across the repetitions of the experiment for each subject and using an order 96 FIR filter implemented in the `eegkit` R-package (Helwig, 2018) to filter the averaged curves at the α frequency bands (between 8 and 12.5 Hz). We thus obtain 64 α-filtered frequency curves for each of the 77 subjects.

Given the low number of observations compared to the 64 functional variables, there is not enough data to reject the null of edge absence even if a true edge were to be present. We therefore aim for a coarser analysis by grouping the variables by brain region and then further according to whether the variable corresponded to the right or left hemispheres of the brain. This yields disjoint groups $G_1, \dots, G_{24}$ comprising 52 variables in total after omitting reference channels and midline channels that could not easily be classified as being in either hemisphere, that is, $G_1 \cup \cdots \cup G_{24} = \{1, \dots, 52\}$. We suppose the observed data are i.i.d. copies functional variables $(X_1, \dots, X_{52})$, and then test the null hypothesis

$$X_{G_j} \perp\!\!\!\perp X_{G_k} \mid \{X_{G_m} : m \in \{1, \dots, 24\} \setminus \{j, k\}\}, \tag{29}$$

for each $j, k \in \{1, \dots, 24\}$ with $j \neq k$; that is, we test for edge presence in the conditional independence graph of the grouped variables. Here, the conditional independence graph over the grouped variables is defined as an undirected graph over $G_1, \dots, G_{24}$, in which the edge between G_j and $G_k, j \neq k$ is missing if and only if (29) holds; that is, rejection of the null in (29) for k and j indicates that the conditional independence graph has an edge between G_k and G_j.

To construct p-values for the null in (29) using the GHCM, we must regress for each $l \in G_j$ and $r \in G_k$, each of the functional variables X_l and X_r on to the set of variables in the conditioning set. Since the regressions will involve large numbers of functional predictors, the `refund` package is not suitable to perform the regressions. Instead, we use the `FDboost` package in `R`, which is well-suited to high-dimensional functional regressions (Brockhaus et al., 2020). We fit a concurrent functional model (Ramsay & Silverman, 2005) of the form

$$X_l(t) = \sum_m \beta_m(t) X_m(t);$$

the inclusion of additional functional linear terms did not improve the fit. We assessed the appropriateness of this regression method to data of the sort studied here through simulations described in Section D of the supplement.

Figure 6 summarises the results of GHCM applied to test the presence of each edge in the conditional independence graph. We see that some of the brain regions located close to each other appear to be connected, as one might expect.

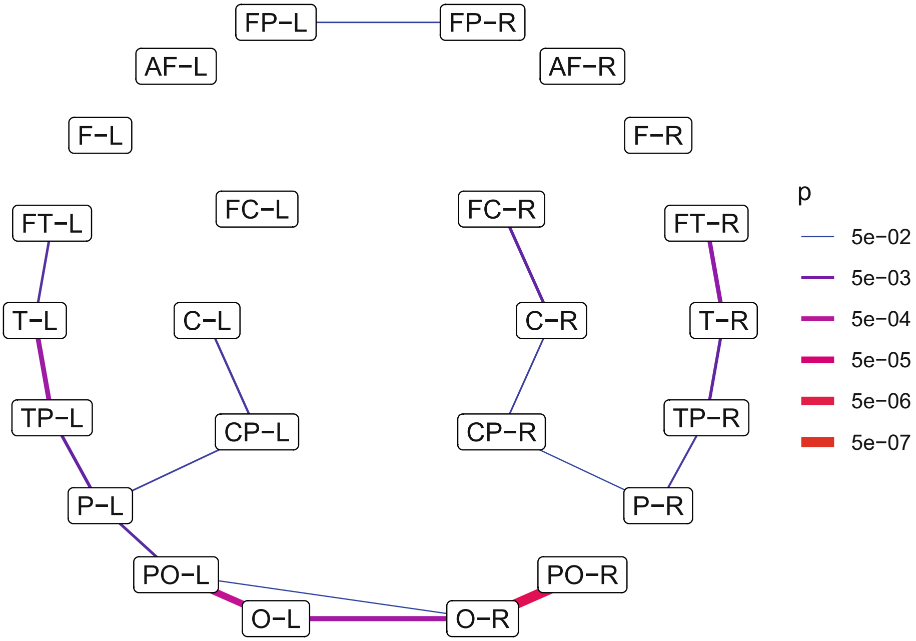

FIGURE 6 Network summarising the output of conditional independence tests for each pair of groups. Only edges with *p*-values of less than 5% are shown with thicker lines indicating smaller *p*-values. [Colour figure can be viewed at wileyonlinelibrary.com]

Note that the network presented includes all edges that had a *p*-value less than 5%. The edge PO-R—O-R has a Bonferroni-corrected *p*-value of 0.0027, and is the only edge yielding a corrected *p*-value less than 5%. Applying the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) to control the false discovery rate at the 5% level selects this edge and also PO-L—O-L. We may compare these results with those of Qiao et al. (2019) and Qiao et al. (2020) who study the same dataset but consider the different problem of estimation of the conditional independence graph rather than testing of edge presence as we do here. We see that our results are broadly in line with their estimates: for example, there are edges estimated between the groups represented by PO-R and O-R (the group pair which yields the lowest *p*-value) even in some of their sparsest estimated graphs.

6 CONCLUSION

Testing the conditional independence $X \perp\!\!\!\perp Y \mid Z$ has been shown to be a hard problem in the setting where X, Y, Z are all real valued and Z is absolutely continuous with respect to Lebesgue measure (Shah & Peters, 2020). This hardness takes a more extreme form in the functional setting: even when (X, Y, Z) are jointly Gaussian with non-degenerate covariance and Z and at most one of X and Y are infinite-dimensional, there is no non-trivial test of conditional independence. This requires us to (i) understand the form of an 'effective null hypothesis' for a given hypothesis test, and (ii) develop tests where these effective nulls are somewhat interpretable so that domain knowledge can more easily inform the choice of a conditional independence test to use on any given dataset.

In order to address these two needs, we introduce here a new family of tests for functional data and develop the necessary uniform convergence results to understand the forms of null hypotheses that we can have type I error control over. We see that for our proposed GHCM tests, error control is guaranteed under conditions largely determined by the in-sample prediction error rate of regressions upon which the test is based. Whilst in-sample and more common out-of-sample results share similarities in some settings, the lack of a need to extrapolate beyond the data in the former lead to important differences when regressing on functional data. In particular, no eigen-spacing conditions or lower bounds on the eigenvalues of the covariance of the regressor are required for the in-sample error to be controlled when ridge regression is used. It would be interesting to investigate the in-sample MSPE properties of other regression methods and understand whether such conditions can be avoided more generally.

One attractive feature of the GHCM is that it only depends on inner products between the residuals produced by the regression methods. An interesting question is whether different inner products can be constructed to have power against different sets of alternatives, by emphasising certain regions of the function domains, for example.

Another direction which may be fruitful to pursue is to adapt the GHCM so that it has power against alternatives where $\mathbb{E}\mathrm{Cov}(X, Y \mid Z) = 0$. It is likely that further conditions will be required of the regression methods than simply that their in-sample prediction errors are small, and so some interpretability of the effective null hypotheses, and indeed its size compared to the full null of conditional independence, will need to be sacrificed. There are however settings where the severity of type I versus type II errors may be balanced such that this is an attractive option.

It would also be interesting to investigate the hardness of conditional independence in the setting where all of X, Y and Z are infinite-dimensional. For our hardness result here, at least one of X and Y must be finite-dimensional. It may be the case that requiring two infinite-dimensional variables to be conditionally independent is such a strong condition that the null is not prohibitively large compared to the entire space of Gaussian measures, and so genuine control of the type I error while maintaining power is in fact possible. Such a result, or indeed a proof that hardness persists, would certainly be of interest.

ACKNOWLEDGEMENTS

We thank Yoav Zemel, Alexander Aue, Sonja Greven and Fabian Scheipl for helpful discussions.

ORCID

Anton Rask Lundborg https://orcid.org/0000-0001-5565-5678
Rajen D. Shah https://orcid.org/0000-0001-9073-3782
Jonas Peters https://orcid.org/0000-0002-1487-7511

REFERENCES

Bai, Z. & Saranadasa, H. (1996) Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6(2), 311–329.

Benatia, D., Carrasco, M. & Florens, J.-P. (2017) Functional linear regression with functional response. *Journal of Econometrics*, 201(2), 269–291.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57(1), 289–300.

Brockhaus, S., Rügamer, D. & Greven, S. (2020) Boosting functional regression models with fdboost. *Journal of Statistical Software*, 94(10), 1–50.

Cai, T.T. & Hall, P. (2006) Prediction in functional linear regression. *Annals of Statistics*, 34(5), 2159–2179.

Chen, X. & White, H. (1998) Central limit and functional central limit theorems for Hilbert-valued dependent heterogeneous arrays with applications. *Econometric Theory*, 14(2), 260–284.

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

Chiou, J.-M., Müller, H.-G. & Wang, J.-L. (2004) Functional response models. *Statistica Sinica*, 14(3), 675–693.

Constantinou, P. & Dawid, A.P. (2017) Extended conditional independence and applications in causal inference. *Annals of Statistics*, 45(6), 2618–2653

Crambes, C. & Mas, A. (2013) Asymptotics of prediction in functional linear regression with functional outputs. *Bernoulli*, 19(5B), 2627–2651.

Delaigle, A. & Hall, P. (2012) Methodology and theory for partial least squares applied to functional data. *Annals of Statistics*, 40(1), 322–352.

Duchesne, P. & de Micheaux, P.L. (2010) Computing the distribution of quadratic forms: further comparisons between the liu-tang-zhang approximation and exact methods. *Computational Statistics and Data Analysis*, 54, 858–862.

Fan, Y., James, G.M. & Radchenko, P. (2015) Functional additive regression. *Annals of Statistics*, 43(5), 2296–2325.

Farebrother, R.W. (1984) Algorithm AS 204: the distribution of a positive linear combination of chi-squared random variables. *Journal of the Royal Statistical Society Series C*, 33(3), 332–339.

Ferraty, F. & Vieu, P. (2006) *Nonparametric functional data analysis: theory and practice*. Springer Series in Statistics. New York: Springer New York.

Ferraty, F., Laksaci, A., Tadj, A. & Vieu, P. (2011) Kernel regression with functional response. *Electronic Journal of Statistics*, 5, 159–171.

Goldsmith, J., Bobb, J., Crainiceanu, C.M., Caffo, B. & Reich, D. (2011) Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20(4), 830–851.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J. et al. (2020) Refund: regression with functional data. R-package version 0.1-22. Available from: https://CRAN.R-project.org/package=refund [Accessed 1st September 2021].

Greven, S. & Scheipl, F. (2017) A general framework for functional regression modelling. *Statistical Modelling*, 17(1–2), 1–35.

Györfi, L., Kohler, M. & Walk, H. (2002) *A distribution-free theory of nonparametric regression*. New York: Springer New York.

Hall, P. & Hooker, G. (2016) Truncated linear models for functional data. *Journal of the Royal Statistical Society Series B*, 78(3), 637–653.

Hall, P. & Horowitz, J.L. (2007) Methodology and convergence rates for functional linear regression. *Annals of Statistics*, 35(1), 70–91.

Helwig, N.E. (2018) eegkit: toolkit for electroencephalography data. R-package version 1.0-4. Available from: https://CRAN.R-project.org/package=eegkit [Accessed 1st September 2021].

Hoerl, A.E. & Kennard, R.W. (2000) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42(1), 80–86.

Imhof, J.P. (1961) Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4), 419–426.

Ingber, L. (1997) Statistical mechanics of neocortical interactions: canonical momenta indicators of electroencephalography. *Physical Review E*, 55, 4578–4593.

Ingber, L. (1998) Statistical mechanics of neocortical interactions: training and testing canonical momenta indicators of EEG. *Mathematical and Computer Modelling*, 27(3), 33–64.

Ivanescu, A.E., Staicu, A.-M., Scheipl, F. & Greven, S. (2015) Penalized function-on-function regression. *Computational Statistics*, 30(2), 539–568.

Koller, D. & Friedman, N. (2009) *Probabilistic graphical models: principles and techniques—adaptive computation and machine learning*. Cambridge, MA: The MIT Press.

Kraft, C. (1955) *Some conditions for consistency and uniform consistency of statistical procedures*. Berkeley, CA: University of California Press.

Lauritzen, S. (1996) *Graphical models*. Oxford Statistical Science Series. Oxford: Clarendon Press.

Liu, H., Tang, Y. & Zhang, H.H. (2009) A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4), 853–856.

Lundborg, A.R., Shah, R.D. & Peters, J. (2022) GHCM: functional conditional independence testing with the GHCM. R-package version 3.0.0. Available from: https://CRAN.R-project.org/package=ghcm [Accessed 1st July 2022].

Morris, J.S. (2015) Functional regression. *Annual Review of Statistics and its Application*, 2(1), 321–359.

Neykov, M., Balakrishnan, S. & Wasserman, L. (2020) Minimax optimal conditional independence testing. *arXiv preprint arXiv:2001.03039*.

Pearl, J. (2009) *Causality*. Cambridge: Cambridge University Press.

Pearl, J. (2014) *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Amsterdam: Elsevier.

Peters, J. (2014) On the intersection property of conditional independence and its application to causal discovery. *Journal of Causal Inference*, 3, 97–108.

Peters, J., Bühlmann, P. & Meinshausen, N. (2016) Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B*, 78(5), 947–1012.

Peters, J., Janzing, D. & Schölkopf, B. (2017) *Elements of causal inference: foundations and learning algorithms*. Cambridge, MA: MIT Press.

Qiao, X., Guo, S. & James, G.M. (2019) Functional graphical models. *Journal of the American Statistical Association*, 114(525), 211–222.

Qiao, X., Qian, C., James, G.M. & Guo, S. (2020) Doubly functional graphical models in high dimensions. *Biometrika*, 107(2), 415–431.

Ramsay, J.O. & Silverman, B.W. (2005) *Functional data analysis*. New York: Springer New York.

Reiss, P.T. & Ogden, R.T. (2007) Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479), 984–996.

Reiss, P.T., Huang, L. & Mennes, M. (2010) Fast function-on-scalar regression with penalized basis expansions. *The International Journal of Biostatistics*, 6(1).

Robins, J.M. & Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429), 122–129.

Scharfstein, D.O., Rotnitzky, A. & Robins, J.M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448), 1096–1120.

Scheipl, F., Staicu, A.-M. & Greven, S. (2015) Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2), 477–501.

Shah, R.D. & Peters, J. (2020) The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3), 1514–1538..

Shin, H. (2009) Partial functional linear regression. *Journal of Statistical Planning and Inference*, 139(10), 3405–3418.

Spirtes, P., Scheines, P., Glymour, C., Scheines, R., Richard, S., Heckerman, D. et al. (2000) *Causation, prediction, and search*. Adaptive computation and machine learning. Cambridge, MA: MIT Press.

Ullah, S. & Finch, C.F. (2013) Applications of functional data analysis: a systematic review. *BMC Medical Research Methodology*, 13(1), 43.

Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016) Functional data analysis. *Annual Review of Statistics and its Application*, 3(1), 257–295.

Wood, S.N. (2013) On *p*-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221–228.

Wood, S.N. (2017) *Generalized additive models*. Boca Raton, FL: Chapman and Hall/CRC.

Yao, F. & Müller, H.-G. (2010) Functional quadratic regression. *Biometrika*, 97(1), 49–64.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005) Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 2873–2903.

Yuan, M. & Cai, T.T. (2010) A reproducing kernel Hilbert space approach to functional linear regression. *Annals of Statistics*, 38(6), 3412–3444.

Yuan, M. & Lin, Y. (2007) Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1), 19–35.

Zapata, J., Oh, S.-Y. & Petersen, A. (2019) Partial separability and functional graphical models for multivariate Gaussian processes. *arXiv preprint arXiv:1910.03134*.

Zhang, X.L., Begleiter, H., Porjesz, B., Wang, W. & Litke, A. (1995) Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38(6), 531–538.

Zhu, H., Strawn, N. & Dunson, D.B. (2016) Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 17(1), 7157–7183.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Lundborg, A.R., Shah, R.D. & Peters, J. (2022) Conditional independence testing in Hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(5), 1821–1850. Available from: https://doi.org/10.1111/rssb.12544