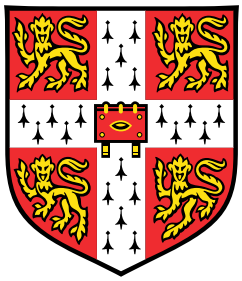
Modern Methods for Variable Significance Testing



Anton Rask Lundborg

Statistical Laboratory
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Hughes Hall

October 2022

In loving memory of my grandparents, Lise and Kalle.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Chapter 2 is joint work with Rajen Shah (University of Cambridge) and Jonas Peters (University of Copenhagen), and has been accepted in the Journal of the Royal Statistical Society: Series B. Chapter 3 is joint work with Ilmun Kim (Yonsei University), Rajen Shah (University of Cambridge) and Richard Samworth (University of Cambridge) and will be submitted for publication soon.

Anton Rask Lundborg
October 2022

Acknowledgements

I would first like to thank my supervisors Rajen Shah and Richard Samworth for their role in bringing me to Cambridge, endless support and generosity with their time. They have taught me much about statistics, the world of academia and how to navigate within it. I would like to thank Rajen for his patience with me during our countless video calls both before, during and after the pandemic – words cannot express how much I have learned about statistical problem-solving from these chats. I would like to thank Richard for constantly inspiring me with his unwavering dedication to academic excellence, I endeavour to continue producing work that lives up to your high standards. A special thank-you goes out to my collaborator and master's supervisor Jonas Peters for encouraging me to pursue a PhD outside my native country. I am not sure that I would have believed I could do it without your supportive words. Thank you to my collaborators Ilmun Kim, Richard Guo and Qingyuan Zhao from whom I have learned much.

I have been extremely lucky to be a part of Richard's group meetings where I had the pleasure of encountering a consistent source of inspiration and new ideas from the many brilliant people attending. I would like to thank Tom Berrett, Tim Cannings, Yudong Chen, Oliver Feng, Bertille Follain, Manuel Müller, Henry Reeve, Tengyao Wang, Min Xu and Yoav Zemel, in particular. My time in the Centre for Mathematical Sciences has been not just intellectually stimulating but downright fun mainly due to the wonderful people there including Joakim Andersen, Tobias Friedling, Harvey Klyne, Ben Stokell, Lennie Wells and Elliot Young. Being able to see and meet up with my friends back home throughout the pandemic was one of the few things keeping me sane through that time, thank you, Adam, Andreas, Stefan and many others from my high school.

I have chosen to dedicate this thesis to the memory of my grandparents who passed away during my time as a PhD student. The influence of their inquisitive minds and kind souls on me cannot be overstated. I am grateful for the support of my mother whose adventurous spirit gave me the final courage needed to move away from home and pursue a PhD. Thank you to my sister for her support of me in all aspects of my life and for being there whenever I felt alone away from home. Thank you to my father and his girlfriend, Lotte, for giving me a place to stay during the many months of lockdown. I can honestly say that I would never have finished the PhD without your help. Finally, I am forever grateful to my father – without your love and support, I would not be who or where I am today.

Abstract

This thesis concerns the ubiquitous statistical problem of *variable significance testing*. The first chapter contains an account of classical approaches to variable significance testing including different perspectives on how to formalise the notion of 'variable significance'. The historical development is contrasted with more recent methods that are adapted to both the scale of modern datasets but also the power of advanced machine learning techniques. This chapter also includes a description of and motivation for the theoretical framework that permeates the rest of the thesis: providing theoretical guarantees that hold uniformly over large classes of distributions.

The second chapter deals with testing the null that $Y \perp\!\!\!\perp X \mid Z$ where X and Y take values in separable Hilbert spaces with a focus on applications to functional data. The first main result of the chapter shows that for functional data it is impossible to construct a non-trivial test for conditional independence even when assuming that the data are jointly Gaussian. A novel regression-based test, called the *Generalised Hilbertian Covariance Measure* (GHCM), is presented and theoretical guarantees for uniform asymptotic Type I error control are provided with the key assumption requiring that the product of the mean squared errors of regressing Y on Z and X on Z converges faster than n^{-1}, where n is the sample size. A power analysis is conducted under the same assumptions to illustrate that the test has uniform power over local alternatives where the expected conditional covariance operator has a Hilbert–Schmidt norm going to 0 at a $\sqrt{n}$-rate. The chapter also contains extensive empirical evidence in the form of simulations demonstrating the validity and power properties of the test. The usefulness of the test is demonstrated by using the GHCM to construct confidence intervals for the boundary point in a truncated functional linear model and to detect edges in a graphical model for an EEG dataset.

The third and final chapter analyses the problem of nonparametric variable significance testing by testing for conditional mean independence, that is, testing the null that $\mathbb{E}(Y \mid X, Z) = \mathbb{E}(Y \mid Z)$ for real-valued Y. A test, called the *Projected Covariance Measure* (PCM), is derived by considering a family of studentised test statistics and choosing a member of this family in a data-driven way that balances robustness and power properties of the resulting test. The test is regression-based and is computed by splitting a set of observations of (X, Y, Z) into two sets of equal size, where one half is used to learn a projection of Y onto X and Z (nonparametrically) and the second half is used to test for vanishing expected conditional correlation given Z between the projection and Y. The chapter contains general conditions that ensure uniform asymptotic Type I control of the resulting test by imposing conditions on the mean-squared error of the involved regressions. A modification of the PCM using additional

sample splitting and employing spline regression is shown to achieve the minimax optimal separation rate between null and alternative under Hölder smoothness assumptions on the regression functions and the conditional density of X given $Z = z$. The chapter also shows through simulation studies that the test maintains the strong type I error control of methods like the Generalised Covariance Measure (GCM) but has power against a broader class of alternatives.

Table of contents

Chapter 1

Introduction

Given a response variable Y and a set of predictors it is a fundamental statistical problem to distinguish between significant (relevant) and insignificant variables. Solving such a problem is crucial to promote the use of parsimonious models for practitioners across scientific fields. In classical statistical modelling, we impose a parametric model on the relationship between response and predictors. Historically, the first such models were linear models (Seal, 1967), where we posit that Y is related to predictors $(X, Z) \in \mathbb{R}^{d_X + d_Z}$ through the relationship

$$Y = \beta^T X + \gamma^\top Z + \varepsilon, \tag{1.1}$$

with $\mathbb{E}(\varepsilon \,|\, X, Z) = 0$. In such a model, there is a clear definition of when X is significant in the presence of Z, namely that $\beta \neq 0$.

In more modern approaches to statistical modelling, we do not directly impose a model on the relationship between response and predictors but instead use highly complex machine learning methods that do not allow as straightforward a definition of significance. These modern algorithms also permit the modelling of increasingly sophisticated data structures such as images, text, audio and curves (Bahdanau et al., 2015; He et al., 2016; Ramsay and Silverman, 2005; Vaswani et al., 2017). For such data it is rarely sensible to model the observations as Euclidean vectors as above but instead a more algorithmic perspective is useful. It is therefore pertinent to move away from a model-based definition of significance (as given for the model in (1.1)) and instead move towards a *model-free* definition. The precise definition of variable significance depends on the context and which sort of inference one is interested in drawing, as we shall investigate shortly.

The primary contribution of this thesis is to provide theoretically sound and practically feasible model-free variable significance testing methods for both conventional and modern data structures.

1.1 Background

1.1.1 Notions of variable significance and previously developed methods

The most well-studied notion of model-free variable significance is that of *conditional independence*. We say that Y is independent of X given Z if $\mathbb{E}(f(Y) \,|\, X, Z) = \mathbb{E}(f(Y) \,|\, Z)$ for any bounded, real-valued and measurable function f. Intuitively, we ask that X provides no additional information for the prediction of any transformation of Y beyond that which is already present in Z. This informal description alludes to the fact that, despite this not being immediately obvious from the definition, conditional independence is a symmetric relation in X and Y (Constantinou and Dawid, 2017, Section 2) and as such we will use $Y \perp\!\!\!\perp X \,|\, Z$ and $X \perp\!\!\!\perp Y \,|\, Z$ interchangeably. Conditional independence is the direct extension of independence to conditional distributions and reduces to independence when Z is constant. It is a convenient simplifying assumption in many cases, perhaps most notably in the increasingly active field of *causal inference*, where conditional independence between certain variables is crucial for obtaining valid inference. Further, using so-called *constraint-based* methods for causal discovery, most famously the PC algorithm (Spirtes et al., 2000) and different variants of Invariant Causal Prediction (ICP) (Peters et al., 2016), it is possible to learn some causal relations from purely observational data by repeatedly testing for conditional independence. In what follows we will always work in the setting of n i.i.d. observations of (X, Y, Z).

When (X, Y, Z) are discrete, a classical family of tests is those of Cochran–Mantel–Haenszel (Cochran, 1954; Mantel, 1963; Mantel and Haenszel, 1959), extending previous work by McNemar (McNemar, 1947). The tests are based on estimating conditional odds ratios and testing whether these equal 1 across each stratum of Z. Another classical example of a conditional independence test is that of the *partial correlation test*. Suppose $(X, Y, Z) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$, and we seek to test whether $X \perp\!\!\!\perp Y \,|\, Z$. Then, we can do a linear (ordinary least squares) regression of X onto Z to yield residuals $\hat{\varepsilon}$ and similarly regress Y onto Z to yield residuals $\hat{\xi}$. The correlation between $\hat{\varepsilon}$ and $\hat{\xi}$ is the *partial correlation* of X and Y given Z. If the vector (X, Y, Z) is Gaussian, then the population partial correlation is zero if and only if $X \perp\!\!\!\perp Y \,|\, Z$ hence the empirical estimate can form the basis of a test. Fisher (1924) derived the distribution of the sample partial correlation under this assumption, which can be used to construct an exact test. If the regression functions for the X on Z and Y on Z regressions are linear (as is the case when the vector (X, Y, Z) is Gaussian), the aforementioned partial correlation test remains an asymptotically valid test, although it is not consistent against all alternatives (Arnold, 1984; Huber, 1973).

The development of a conditional independence test under weaker assumptions than those given above is a difficult task as illustrated by the main result of Shah and Peters (2020). In this work, the authors show that no test of conditional independence of size α over a class of distributions containing those where (X, Y, Z) has a density with respect to Lebesgue measure can have power greater than α against any alternative. In other words, no non-trivial conditional independence test exists without shrinking the class of null distributions from the full nonparametric model. One conclusion to be drawn from this result is the importance of transparent assumptions when developing new conditional independence tests. It is essential

that we choose a suitable test for any given set of data and encoding the assumptions in such a way that we are able to select a test is therefore of the utmost importance.

One way to gauge the suitability of a test to a given application is through describing which (hopefully large) class of null distributions that we can expect the test to be (asymptotically) valid over. The partial correlation test is valid over the set of Gaussian distributions and asymptotically valid over classes with linear regression functions and tails that are not too heavy, both subclasses of the set of distributions that are absolutely continuous with respect to Lebesgue measure. Shah and Peters (2020) propose the Generalised Covariance Measure (GCM) as a generalisation of the partial correlation test to non-linear settings, that permits validity over larger classes still. Using arbitrary regression methods we regress X onto Z yielding residuals $\hat{\varepsilon}$ and Y onto Z yielding residuals $\hat{\xi}$ and set $R_i := \hat{\varepsilon}_i \hat{\xi}_i$. The nominal level $\alpha \in (0, 1)$ GCM rejects the null hypothesis of conditional independence whenever the absolute value of

$$T = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} R_i}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} R_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} R_i\right)^2}}$$

exceeds the $(1 - \alpha/2)$ quantile $z_{1-\alpha/2}$ of the standard Gaussian distribution. The GCM is asymptotically valid under mild conditions, of which the most restrictive is that the product of the mean-squared errors of the aforementioned regressions goes to 0 at a rate of n^{-1}. This includes cases where one of the regression is parametric and the other is merely consistent, and cases where the regression functions are both moderately smooth. We see that the restriction of the class of null distributions is done through the selection of appropriate regression methods. Selecting a suitable regression method is a common statistical task and practitioners can therefore leverage their intuition and experience with this task when using the GCM.

Another modern example of a conditional independence test, coming from the machine learning literature, is the Kernel Conditional Independence test (KCI) (Zhang et al., 2011). The test is similar in spirit to the GCM in that the final test is based on a measure of correlation between residuals. However, before doing any regressions, the variables are embedded into a Reproducing Kernel Hilbert Space (RKHS) with the goal of detecting non-linear departures from the null (in contrast to the partial correlation test) and kernel ridge regression is used as the regression method. To compute the test of $X \perp\!\!\!\perp Y \mid Z$, the vector (X, Z) is embedded into an RKHS and kernel ridge regression is used to regress out the Z part of this embedding. Similarly, Y is embedded into an RKHS and Z is regressed out. The resulting residuals are computed efficiently using RKHS theory and an asymptotic distribution of a correlation measure can be derived under the null. The KCI, like the GCM, also restricts the class of nulls by means of the validity of a regression method, namely kernel ridge regression. The KCI is highly popular especially among the computer science community, however our experience is that it is difficult to choose the tuning parameters of the test effectively, see Section 3.6.

There are other ways of shrinking the class of null distributions from the full nonparametric model than by requiring powerful regression methods. In the case that one has complete knowledge of one of the conditional distributions, it turns out that it is now possible to construct correctly calibrated tests with non-trivial power. This assumption, often referred

to as the 'model-X' assumption, has seen increasing attention in the recent years, perhaps most notably by Candès et al. (2018) and Berrett et al. (2020). The latter paper and Barber et al. (2020) investigate the use of an estimate of the conditional distribution in the 'model-X' framework. The analysis reveals that the estimate needs to be of high quality for the validity of the method to be guaranteed.

Canonne et al. (2018) take a minimax perspective on conditional independence testing for discrete data. Neykov et al. (2021) and later Kim et al. (2021) extend this to continuous data by a family of tests based on binning the observations into bins with size varying with n. The authors proceed to show that their tests are minimax optimal under certain smoothness conditions on the conditional density of (X, Y) given Z and when this smoothness level is known.

Few of the methods mentioned above have immediate analogues for non-Euclidean data structures such as functional data. Functional Data Analysis (FDA) traces its roots to the work of Grenander (1950) and Rao (1958) before becoming an active field around the turn of the century through the work of Ramsay, Silverman, Rice and others (Ramsay, 1982; Rice and Silverman, 1991). In recent years, more work has been done on functional regression methods (Goldsmith et al., 2011; Ivanescu et al., 2015; Morris, 2015; Reiss et al., 2016), extending generalised linear or additive models to the setting of scalar-on-function or function-on-function regression. Several software packages have also been developed for these methods including `FDboost` (Brockhaus et al., 2020) and `refund` (Goldsmith et al., 2020). The `refund` package utilises the extensive libraries on generalised additive models (see Wood (2017)) to fit a variety of functional regression methods. For generalised additive models, it is possible to derive heuristic p-values for the inclusion of a predictor Wood (2013) and these are also reported in the functional data package. Extending the GCM as described above to the setting of functional data is the subject of Chapter 2, providing a practical test with interpretable conditions ensuring the validity of the procedure.

In our discussion thus far we have focused on variable significance defined through conditional independence but when the goal of the modelling is to merely predict Y, a natural condition is instead that of *conditional mean independence*. We say that Y is conditionally mean independent of X given Z if $\mathbb{E}(Y \mid X, Z) = \mathbb{E}(Y \mid Z)$. In words, we are asking that the best predictor of Y given X and Z equals the best predictor given just Z. This is a non-symmetric condition – it is perfectly possible that X is a significant predictor for Y but Y is an insignificant predictor of X, e.g. when $Y = X^2 + \varepsilon$ and X and ε are independent standard Gaussian. Comparatively little work has been done on testing this null (Aït-Sahalia et al., 2001; Fan and Li, 1996; Lavergne and Vuong, 2000).

It turns out that the GCM is in fact also a test of conditional mean independence as the GCM is only powerful when

$$\mathbb{E}(\mathrm{Cov}(X, Y \mid Z)) := \mathbb{E}(\{X - \mathbb{E}(X \mid Z)\}\{Y - \mathbb{E}(Y \mid Z)\}) \neq 0,$$

and this is never true under conditional mean independence. However, there are cases where conditional mean independence is violated (and therefore also conditional independence) but $\mathbb{E}(\mathrm{Cov}(X, Y \mid Z)) = 0$, e.g. the example given in the previous paragraph with Z independent of

(X, ε). Testing for conditional mean independence with optimal power properties is the subject of Chapter 3.

1.1.2 Modern theory of hypothesis testing

Many classical (and even some more modern) papers on hypothesis testing follow a simple structure; develop a test statistic T_n, show that it converges in distribution to some known limit (typically a standard Gaussian or χ^2_1-distribution) and construct a test that rejects based on quantiles of the limiting distribution. Let $\mathcal{P}_0$ denote the class of null distributions for which we wish to control Type I error, that is, we seek

$$\sup_{P \in \mathcal{P}_0} \mathbb{P}_P(\text{reject } H_0) \leq \alpha \tag{1.2}$$

for a user-specified significance level $\alpha \in (0, 1)$. A test satisfying (1.2) is said to be *level* α. The approach outlined above instead shows that

$$\sup_{P \in \mathcal{P}_0} \lim_{n \to \infty} \mathbb{P}_P(\text{reject } H_0) \leq \alpha. \tag{1.3}$$

A test satisfying (1.3) is said to be *pointwise asymptotic level* α. This latter property turns out to be a rather weak condition – tests that are pointwise asymptotic level α can have exact level 1 for all n as the following example, due to Romano (2004), illustrates. Consider the problem of testing whether univariate X is mean zero when assuming that X has finite variance. The t-test is clearly pointwise asymptotic level by the CLT. However, consider the distribution that puts mass $(1-p)$ on p and p on $-(1-p)$ for $p \in (0, 1)$. Consider the observation $x = (\overbrace{p, \ldots, p}^{n})$ and note that the probability of observing this x is $(1-p)^n$. Since such an observation would imply rejection of the null using the t-test at any significance level, we have that

$$\mathbb{P}(\text{reject } H_0) \geq (1-p)^n.$$

Thus, choosing p going to 0 yields that the size is 1. Worse still, the Bahadur–Savage theorem (Bahadur and Savage, 1956) shows that no non-trivial test of this null exists.

The approach of both remaining chapters in this thesis, is to find tests satisfying

$$\lim_{n \to \infty} \sup_{P \in \mathcal{P}_0} \mathbb{P}_P(\text{reject } H_0) \leq \alpha. \tag{1.4}$$

Tests satisfying (1.4) are said to *uniformly asymptotically level* α. This requirement is clearly stronger than (1.3) and examples like the one above for the t-test are not possible for uniformly asymptotically level tests. Showing that tests satisfy (1.4) requires a calculus of uniform stochastic convergence that is developed throughout the thesis.

For tests that are correctly calibrated under the null (i.e. reject with probability no more than the chosen significance level α), it is pertinent to ask about the optimality of a given test in terms of the ability of the test to reject false hypotheses. The *minimax* paradigm is one measure of optimality that is gaining in popularity in recent years (Balakrishnan and

Wasserman, 2018; Baraud, 2002; Berrett et al., 2021; Diakonikolas and Kane, 2016; Valiant and Valiant, 2017). In the context of nonparametric hypothesis testing this framework is typically attributed to Ingster (Ingster, 1982, 1987). Let $\mathcal{P}$ denote a family of distributions and partition $\mathcal{P}$ into $\mathcal{P}_0$ (null distributions) and $\mathcal{P}_1$ (alternative distributions). Let d denote a metric on $\mathcal{P}$ and let $\mathcal{P}_1(\epsilon) := \{P \in \mathcal{P}_1 : d(\mathcal{P}_0, P) \geq \epsilon\}$. We seek to characterise the fastest rate at which we can let ϵ go to zero as a function of n while still ensuring that

$$R(\epsilon) := \sup_{P \in \mathcal{P}_0} \mathbb{P}_P(\text{reject } H_0) + \sup_{P \in \mathcal{P}_1(\epsilon)} \mathbb{P}_P(\text{accept } H_0)$$

remains bounded above by a small constant for a given test under consideration. This is a *minimax upper bound* for the radius ϵ using the test. If the upper bound matches a lower bound taken over all possible tests in terms of how it scales in the sample size n, the test is *minimax optimal*. This implies that the test stays consistent with respect to alternative distributions where $d(\mathcal{P}_0, P)$ goes to 0 at the fastest possible rate where testing is possible.

In Chapter 3, one of the main goals is to characterise the critical radius ϵ_n for testing conditional mean independence under smoothness conditions.

1.2 Notation

While some notation is consistent throughout the remaining chapters, there will be many conventions that are local to each chapter. Notation established in each chapter should therefore be considered specific to the chapter where it is introduced.

Chapter 2

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

2.1 Introduction

In a variety of application areas, such as meteorology, neuroscience, linguistics, and chemometrics, we observe samples containing random functions (Ramsay and Silverman, 2005; Ullah and 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 and Hall, 2012; Goldsmith et al., 2011; Hall and Horowitz, 2007; Reiss and Ogden, 2007; Shin, 2009; Yuan and 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 et al., 2011; Ferraty and Vieu, 2006; Yao and Müller, 2010), and versions able to handle potentially large numbers of functional predictors (Fan et al., 2015), to give a few examples; see Morris (2015); Wang et al. (2016) 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 2.1.2. 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 and Friedman, 2009; Lauritzen, 1996) and variable selection. For example, consider the linear scalar-on-function regression model

$$Y = \int_0^1 \theta_X(t)X(t)dt + \int_0^1 \theta_Z(t)Z(t)dt + \varepsilon, \tag{2.1}$$

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 $X \perp\!\!\!\perp Y \,|\, Z$ is equivalent to $\theta_X = 0$, i.e., whether the functional predictor X is significant.

For nonlinear regression models, the conditional independence $X \perp\!\!\!\perp Y \,|\, Z$ 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 $X_1, \ldots, X_p$ are predictors, some or all of which may be functional. Then a set of predictors $S \subseteq \{1, \ldots, p\}$ that contain all useful information for predicting Y, that is such that $Y \perp\!\!\!\perp \{X_j\}_{j \notin S} \,|\, \{X_j\}_{j \in S}$, is known as a Markov blanket of Y in the graphical modelling literature (Pearl, 2014, Sec. 3.2.1). If $Y \not\perp\!\!\!\perp X_j \,|\, \{X_k\}_{k \neq j}$, 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 and 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 $X \perp\!\!\!\perp Y \,|\, Z$ 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., 2021).

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 (2.1), testing for $X \perp\!\!\!\perp Y \,|\, Z$ 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 and 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 and 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, whilst being computationally efficient, however do not have formal uniform level control guarantees.

2.1.1 Our main contributions and organisation of the chapter

It is impossible to test conditional independence with Gaussian functional data.

In Section 2.2 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 2.7. 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.

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 2.8 which studies uniform convergence of Hilbertian and Banachian random variables.

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 2.3 we describe our new GHCM testing framework for testing $X \perp\!\!\!\perp Y \,|\, Z$, where each of X, Y and Z may be collections of functional and scalar variables. In Section 2.4 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 $\mathbb{E}(X \,|\, Z)$ 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, Thm. 3.1). Thus, the regression methods must be chosen appropriately in order for the GHCM to perform well. In Section 2.4.3 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.

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. Whilst 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 2.4.4 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 and Hall, 2006; Crambes and Mas, 2013; Hall and Horowitz, 2007).

The GHCM has several uses.

Section 2.5 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 2.1). (ii) Deriving confidence intervals for truncation points in truncated functional linear model. We demonstrate in Section 2.5.2 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 2.5.3, we use the GHCM to learn functional graphical models for EEG data from a study on alcoholism.

We conclude with a discussion in Section 2.6 outlining potential follow-on work and open problems. The remaining sections contain 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., 2021b) implementing the methodology is available on CRAN.

2.1.2 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 $X \perp\!\!\!\perp Y \,|\, Z$ when

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

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 $X \perp\!\!\!\perp Y \,|\, Z$ is that 'knowing Z renders X irrelevant for predicting Y' (Lauritzen, 1996).

Throughout the chapter 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 $X \perp\!\!\!\perp Y \,|\, Z$, 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 this data a test $\psi_n : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^n \to \{0, 1\}$, with a value of 1 indicating rejection. At times, we will 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 (see Section 2.3.2), $X^{(n)}$ 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 $\mathscr{A}$ on $\mathcal{H}$ is compact if it has a singular value decomposition, i.e., 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

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

for all $h \in \mathcal{H}$. For a compact linear operator $\mathcal{A}$ as above, we denote by $\|\mathscr{A}\|_{\text{op}}$, $\|\mathscr{A}\|_{\text{HS}}$ and $\|\mathscr{A}\|_{\text{TR}}$ the operator norm, Hilbert–Schmidt norm and trace norm, respectively, of $\mathscr{A}$, 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

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

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

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

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

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

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.

2.2 Hardness of 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 (2.1) 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 2.7 in Section 2.2); 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 2.1 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 2.1. *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 2.1 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 2.1 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 $\mathcal{P}_0^Q$. It should be noted that this overall argument is similar to the proof of the hardness result in Shah and Peters (2020) but the exact construction of the mixture of nulls is entirely distinct and is more closely related to constructions used in the proofs of minimax lower bounds in functional estimation.

A consequence of Theorem 2.1 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 chapter is devoted to developing a family of conditional independence tests whose validity rests primarily on the prediction errors of these regressions.

2.3 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 2.3.2 we then define the GHCM.

2.3.1 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 $f(z) := \mathbb{E}(X \,|\, Z = z)$ and similarly $Y = g(Z) + \xi$ with $g(z) := \mathbb{E}(Y \,|\, Z = z)$. The conditional covariance of X and Y given Z,

$$\mathrm{Cov}(X, Y \,|\, Z) := \mathbb{E}\left[\{X - \mathbb{E}(X \,|\, Z)\}\{Y - \mathbb{E}(Y \,|\, Z)\} \,|\, Z\right] = \mathbb{E}(\varepsilon\xi \,|\, Z),$$

has the property that $\mathrm{Cov}(X, Y \,|\, Z) = 0$ and hence $\mathbb{E}(\varepsilon\xi) = 0$ whenever $X \perp\!\!\!\perp Y \,|\, Z$. 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

$i = 1, \dots, n$ and then averaged to give $\bar{R} := \sum_{i=1}^n R_i/n$, an estimate of $\mathbb{E}(\varepsilon\xi)$. The standard deviation of $\bar{R}$ under the null $X \perp\!\!\!\perp Y \mid Z$ may also be estimated, and it can be shown (Shah and Peters, 2020, Thm 8) that under some conditions, $\bar{R}$ 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,

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

This is a zero matrix when $X \perp\!\!\!\perp Y \mid Z$, and hence $\mathbb{E}(\varepsilon\xi^\top) = 0$ under this null. Thus, $\bar{R}$ 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 $\bar{R} \in \mathbb{R}^{d_X \times d_Y}$. The approach taken in Shah and Peters (2020) involves taking the maximum absolute value of a version of $\bar{R}$ 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 $\bar{R}$ by $\sqrt{n}$:

$$\begin{aligned}
\sqrt{n}\bar{R} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\varepsilon}_i\hat{\xi}_i^\top = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(z_i) - \hat{f}(z_i) + \varepsilon_i)(g(z_i) - \hat{g}(z_i) + \xi_i)^\top \\
&= \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i\xi_i^\top}_{U_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(z_i) - \hat{f}(z_i))(g(z_i) - \hat{g}(z_i))^\top}_{a_n} \\
&\quad + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(z_i) - \hat{f}(z_i))\xi_i^\top}_{b_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i(g(z_i) - \hat{g}(z_i))^\top}_{c_n}.
\end{aligned} \tag{2.2}$$

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 $\|\cdot\|_F$ to the a_n term, we get by submultiplicativity and the Cauchy–Schwarz inequality,

$$\begin{aligned}
\|a_n\|_F &\leq \frac{1}{\sqrt{n}}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\|_2 \|g(z_i) - \hat{g}(z_i)\|_2 \\
&\leq \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\|_2^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n \|g(z_i) - \hat{g}(z_i)\|_2^2\right)^{1/2},
\end{aligned} \tag{2.3}$$

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 $\sqrt{n}\bar{R}$ 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 (2.3) 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, $\sqrt{n}\bar{R}$ 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 $t(\sqrt{n}\bar{R})$ 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 $\bar{R}$, i.e., view the matrix as a $d_X d_Y$-dimensional vector, then under the assumptions required for the above heuristic arguments to formally hold, $\sqrt{n}\mathrm{Vec}(\bar{R})$ 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 $X \perp\!\!\!\perp Y \,|\, Z$. Provided C is invertible, $\sqrt{n}C^{-1/2}\bar{R}$ therefore converges to a Gaussian with identity covariance under the null and hence $\|C^{-1/2}\sqrt{n}\bar{R}\|_2^2$ 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.

2.3.2 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. It should be noted that the finite-dimensional setting does not necessitate a functional perspective on the data but if the data is functional, then regression methods respecting this should be superior in terms of performance and the asymptotics of the finite-dimensional case could include 'fill-in asymptotics', where the number of observations per curve increases with n.

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 2.3.2 we set out a scheme to handle this case and bring it within our framework here.

Similarly to the approach outlined in Section 2.3.1, 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 $i = 1, \ldots, n$,

$$\mathscr{R}_i := \hat{\varepsilon}_i \otimes \hat{\xi}_i, \quad \text{and} \quad \mathscr{T}_n := \sqrt{n}\bar{\mathscr{R}} \tag{2.4}$$
$$\text{where} \quad \bar{\mathscr{R}} := \frac{1}{n}\sum_{i=1}^{n} \mathscr{R}_i.$$

We can show (see Theorem 2.2) that under the null, provided the analogous prediction error terms in (2.3) decay sufficiently fast and additional regularity conditions hold, $\mathscr{T}_n$ 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

$$\hat{\mathscr{C}} := \frac{1}{n-1}\sum_{i=1}^{n}(\mathscr{R}_i - \bar{\mathscr{R}}) \otimes_{\text{HS}} (\mathscr{R}_i - \bar{\mathscr{R}}) \tag{2.5}$$

where $\otimes_{\text{HS}}$ 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 $\hat{\mathscr{C}}$. However, here we hit a clear obstacle: even in the finite-dimensional setting, whenever $d_X d_Y \geq n$, the inverse of $\hat{\mathscr{C}}$ 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

$$T_n := \|\mathscr{T}_n\|_{\text{HS}}^2 \tag{2.6}$$

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

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\hat{\varepsilon}_i, \hat{\varepsilon}_j\rangle\langle\hat{\xi}_i, \hat{\xi}_j\rangle; \tag{2.7}$$

see Section 2.9.1 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 $\mathscr{T}_n$ 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 $\mathscr{T}_n$. This leads to the α-level test function ψ_n given by

$$\psi_n := \mathbb{1}_{\{T_n \geq q_\alpha\}} \tag{2.8}$$

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 $\hat{\mathscr{C}}$. 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

$$\Gamma_{ij} := \langle \hat{\varepsilon}_i, \hat{\varepsilon}_j \rangle \langle \hat{\xi}_i, \hat{\xi}_j \rangle; \tag{2.9}$$

see Section 2.9.1 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 and 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, i.e., 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

$$\hat{\varepsilon}_i^\top \hat{\varepsilon}_j = \sum_{k=1}^{d_X} W_{\hat{\varepsilon},i}(k/d_X) W_{\hat{\varepsilon},i}(k/d_X) \qquad \text{and} \qquad \int_0^1 W_{\hat{\varepsilon},i}(t) W_{\hat{\varepsilon},j}(t)\,\mathrm{d}t. \tag{2.10}$$

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 2.4 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 (2.10), and an inner product that maintains this approximation may be more desirable from a power perspective.

Algorithm 1 Generalised Hilbertian Covariance Measure (GHCM)

input: $X^{(n)} \in \mathbb{R}^{n \times d_X}$, $Y^{(n)} \in \mathbb{R}^{n \times d_Y}$, $Z^{(n)} \in \mathbb{R}^{n \times d_Z}$.
options: regression methods for each of the regressions.

1: regress $X^{(n)}$ on $Z^{(n)}$ producing residuals $\hat{\varepsilon}_i \in \mathbb{R}^{d_X}$ for $i = 1, \dots, n$.
2: regress $Y^{(n)}$ on $Z^{(n)}$ producing residuals $\hat{\xi}_i \in \mathbb{R}^{d_Y}$ for $i = 1, \dots, n$.
3: 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 (2.9)).
4: compute test statistic $T_n \leftarrow \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \Gamma_{ij}$.
5: 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 $1/n$.
6: compute the non-zero eigenvalues $\lambda_1, \dots, \lambda_d$ of A (there are at most $n - 1$).
7: 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.

output: p-value p.

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., 2021b) implementing the methodology is available on CRAN.

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.

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 (2.9) which we may then use in lines 7 and following in

Algorithm 1. Furthermore, Theorems 2.3 and 2.4 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.

2.4 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.

2.4.1 Background on uniform convergence

In Section 2.2 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 2.1. *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 $(X, Y) \perp\!\!\!\perp (Z_{r+1}, Z_{r+2}, \ldots) \,|\, Z_1, \ldots, Z_r$ for some $r \in \mathbb{N}$ and $Z_k := \langle e_k, Z\rangle$ for all $k \in \mathbb{N}$. Fix $Q \in \mathcal{Q}$ and recall the definition of $\mathcal{P}_0^Q$ from Section 2.2. Then, for any test ψ_n,*

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

In other words, even if we know a basis $(e_k)_{k\in\mathbb{N}}$ such that in particular the conditional expectations $\mathbb{E}(X \,|\, Z)$ and $\mathbb{E}(Y \,|\, Z)$ are sparse in that they depend only on finitely many components $Z_1, \ldots, Z_r$ (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 $\mathcal{Q}$ 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 $Y, Z_1, \ldots, Z_{a(n)}$ 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,

$$\sup_{P\in\mathcal{P}_0} \lim_{n\to\infty} \mathbb{P}_P(\psi_n^{\mathrm{OLS}} = 1) = \alpha \qquad \text{and} \qquad \inf_{Q\in\mathcal{Q}} \lim_{n\to\infty} \mathbb{P}_Q(\psi_n^{\mathrm{OLS}} = 1) = 1; \tag{2.11}$$

see Section 2.9.2 for a derivation. This illustrates the difference between pointwise asymptotic level control in the left-hand side of (2.11), 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

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

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

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

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 2.8. Throughout this section we emphasise the dependence of many of the quantities in Section 2.3.1 on the distribution of (X, Y, Z) with a subscript P, e.g. f_P, ε_P etc.

In Sections 2.4.2 and 2.4.3 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 2.4.4 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.

2.4.2 Size of the test

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

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

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

$$M^f_{n,P} := \frac{1}{n}\sum_{i=1}^n \left\|f_P(z_i) - \hat{f}^{(n)}(z_i)\right\|^2, \qquad M^g_{n,P} := \frac{1}{n}\sum_{i=1}^n \left\|g_P(z_i) - \hat{g}^{(n)}(z_i)\right\|^2, \tag{2.12}$$

$$\tilde{M}^f_{n,P} := \frac{1}{n}\sum_{i=1}^n \left\|f_P(z_i) - \hat{f}^{(n)}(z_i)\right\|^2 v_P(z_i), \qquad \tilde{M}^g_{n,P} := \frac{1}{n}\sum_{i=1}^n \left\|g_P(z_i) - \hat{g}^{(n)}(z_i)\right\|^2 u_P(z_i). \tag{2.13}$$

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 (2.12) being small, the operator-valued statistic $\mathscr{T}_n$ converges in distribution uniformly to a mean zero Gaussian whose covariance can be estimated consistently. We remark that prediction error quantities in (2.12) and (2.13) 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 2.2. *Let $\tilde{\mathcal{P}}_0 \subseteq \mathcal{P}_0$ be such that uniformly over $\tilde{\mathcal{P}}_0$,*

(i) $nM^f_{n,P}M^g_{n,P} \overset{P}{\rightrightarrows} 0$,

(ii) $\tilde{M}^f_{n,P} \overset{P}{\rightrightarrows} 0$, $\tilde{M}^g_{n,P} \overset{P}{\rightrightarrows} 0$,

(iii) $\inf_{P\in\tilde{\mathcal{P}}_0} \mathbb{E}_P\left(\|\varepsilon_P\|^2\|\xi_P\|^2\right) > 0$ *and* $\sup_{P\in\tilde{\mathcal{P}}_0} \mathbb{E}_P\left(\|\varepsilon_P\|^{2+\eta}\|\xi_P\|^{2+\eta}\right) < \infty$ *for some* $\eta > 0$, *and*

(iv) 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

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

where

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

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 and 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 $M^f_{n,P}$, $M^g_{n,P} \overset{P}{\rightrightarrows} 0$. Many of the aforementioned methods either impose Donsker conditions or employ sample-splitting or cross-fitting as an essential tool to mitigate bias, however, for our result above these conditions are not required. This is in contrast to the power analysis in next section where such conditions cannot be avoided.s We note that importantly, the regression methods are not required to extrapolate well beyond the observed data. We show in Section 2.4.4 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, Lem. 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 (2.8) has type I error control uniformly over $\tilde{\mathcal{P}}_0$ given in Theorem 2.2, provided an additional assumption of non-degeneracy of the covariance operators is satisfied.

Theorem 2.3. *Let $\tilde{\mathcal{P}}_0 \subseteq \mathcal{P}_0$ satisfy the conditions stated in Theorem 2.2, and in addition suppose*

$$\inf_{P \in \tilde{\mathcal{P}}_0} \|\mathscr{C}_P\|_{op} > 0. \tag{2.14}$$

Then for each $\alpha \in (0,1)$, the α-level GHCM test ψ_n (2.8) satisfies

$$\lim_{n\to\infty} \sup_{P \in \tilde{\mathcal{P}}_0} |\mathbb{P}_P(\psi_n = 1) - \alpha| = 0. \tag{2.15}$$

2.4.3 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 have yet to see a concrete example where sample-splitting was required for power. As we require $\hat{f}$ and $\hat{g}$ to satisfy (i) and (ii) of Theorem 2.2, 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 2.4. *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 2.2 and (2.14) with $\tilde{\mathcal{P}}$ in place of $\tilde{\mathcal{P}}_0$. Then writing $\mathscr{K}_P := \mathbb{E}_P(\varepsilon_P \otimes \xi_P) = \mathbb{E}_P(\mathrm{Cov}_P(X, Y \mid Z))$, we have, uniformly over $\tilde{\mathcal{P}}$,*

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

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

(i) Redefining $\tilde{\mathcal{P}}_0 = \tilde{\mathcal{P}} \cap \mathcal{P}_0$, we have that (2.15) is satisfied, and so an α-level GHCM test has size converging to α uniformly over $\tilde{\mathcal{P}}_0$.

(ii) For every $0 < \alpha < \beta < 1$ there exists $c > 0$ and $N \in \mathbb{N}$ such that for any $n \geq N$,

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

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

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

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

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

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

where $v(t,u)$ is 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 2.5.1.

2.4.4 GHCM using linear function-on-function ridge regression

Here we consider a special case of the general setup used in Sections 2.4.2 and 2.4.3 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:

$$X = \mathscr{S}_P^X Z + \varepsilon_P \tag{2.16}$$

$$Y = \mathscr{S}_P^Y Z + \xi_P. \tag{2.17}$$

Here $\mathscr{S}_P^X$ is a Hilbert–Schmidt operator such that $\mathscr{S}_P^X Z = f(Z) := \mathbb{E}(X \mid Z)$, with analogous properties holding for $\mathscr{S}_P^Y$, 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

$$X(t) = \int_0^1 \beta_P^X(s,t) Z(s) \, \mathrm{d}s + \varepsilon_P(t), \tag{2.18}$$

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,

Chap. 16), and studied by Chiou et al. (2004); Crambes and Mas (2013); Yao et al. (2005), for example. Benatia et al. (2017) propose a Tikhonov regularised estimator analogous to ridge regression (Hoerl and Kennard, 2000); applied to the regression model (2.16), this estimator takes the form

$$\hat{\mathscr{S}} = \operatorname*{argmin}_{\mathscr{S}} \sum_{i=1}^{n} \|x_i - \mathscr{S}(z_i)\|^2 + \gamma \|\mathscr{S}\|_{\mathrm{HS}}^2, \tag{2.19}$$

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

We now consider a specific instance of the general GHCM framework using regression estimates based on (2.19). Specifically, we form estimate $\hat{\mathscr{S}}^X$ of $\mathscr{S}^X$ by solving the optimisation in (2.19) with regularisation parameter

$$\hat{\gamma} := \operatorname*{argmin}_{\gamma > 0} \left(\frac{1}{\gamma n} \sum_{i=1}^{n} \min(\hat{\mu}_i/4, \gamma) + \frac{\gamma}{4} \right), \tag{2.20}$$

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 $\hat{\mathscr{S}}^Y$ of $\mathscr{S}^Y$ analogously but with the x_i replaced by y_i in (2.19). Note that in the case where $K = 0$ and so $\hat{\gamma}$ does not exist, we simply take $\hat{\mathscr{S}}^X$ and $\hat{\mathscr{S}}^Y$ to be 0 operators, i.e., 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 $\hat{\mathscr{S}}^X$ and $\hat{\mathscr{S}}^Y$ (see Lemma 2.17) where we have omitted some distribution-dependent factors of $\|\mathscr{S}_P^X\|_{\mathrm{HS}}^2$ or $\|\mathscr{S}_P^Y\|_{\mathrm{HS}}^2$ and a variance factor; a similar strategy was used in an analysis of kernel ridge regression (Shah and 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 (2.8) with estimated regression functions $\hat{f}$ and $\hat{g}$ yielding fitted values given by

$$\hat{f}(z_i) = \hat{\mathscr{S}}^X z_i \qquad \text{and} \qquad \hat{g}(z_i) = \hat{\mathscr{S}}^Y z_i, \qquad \text{for all } i = 1, \ldots, n. \tag{2.21}$$

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 2.5. *Let $\tilde{\mathcal{P}}_0 \subset \mathcal{P}_0$ be such that (2.16) and (2.17) are satisfied, and moreover (iii) and (iv) of Theorem 2.2 and (2.14) hold when $\hat{f}$ and $\hat{g}$ are as in (2.21). Suppose further that*

(i) $\sup_{P \in \tilde{\mathcal{P}}_0} \max(\|\mathscr{S}_P^X\|_{HS}, \|\mathscr{S}_P^Y\|_{HS}) < \infty$,

(ii) $\sup_{P \in \tilde{\mathcal{P}}_0} \max(u_P(Z), v_P(Z)) < \infty$ *almost surely,*

(iii) $\sup_{P \in \tilde{\mathcal{P}}_0} \mathbb{E}\|Z\|^2 < \infty$ *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 2.5 are satisfied, we may consider the family of distribution $\tilde{\mathcal{P}}_0$ where Z, ε_P in (2.22) and ξ_P in (2.23) are independent, and the latter two are Brownian motions with variances $\sigma^2_{\varepsilon,P}$ and $\sigma^2_{\xi,P}$ respectively. If the coefficient functions β^X_P corresponding to X in (2.18) 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 $\sigma^2_{\varepsilon,P}$ and $\sigma^2_{\xi,P}$ are bounded from above and below uniformly, we have that $\mathcal{P}_0$ satisfies all the requirements of Theorem 2.5.

The proof of Theorem 2.5 relies on Lemma 2.17 in Section 2.9.5, 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 2.2). 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 and Hall, 2006; Hall and 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.

2.5 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 2.5.3 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 2.4.4) 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 2.10, 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.

2.5.1 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 2.5.1 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. In what follows we do a modest number of simulations to get a broad overview of the behaviour of our test in a variety of settings. Each plot produced here is the result of between 80-120 computing hours on dedicated machines.

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

$$X(t) = \int_0^1 \beta_a(s,t) Z(s) \, \mathrm{d}s + N_X(t), \tag{2.22}$$

$$Y = \int_0^1 \alpha_a(t) Z(t) \, \mathrm{d}t + N_Y. \tag{2.23}$$

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

$$\beta_a(s,t) = a \exp(-(st)^2/2) \sin(ast), \qquad \alpha_a(t) = \int_0^1 \beta_a(s,t) \, \mathrm{d}s. \tag{2.24}$$

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 (2.22), (2.23), 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 2.1. We see that the `pfr` 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

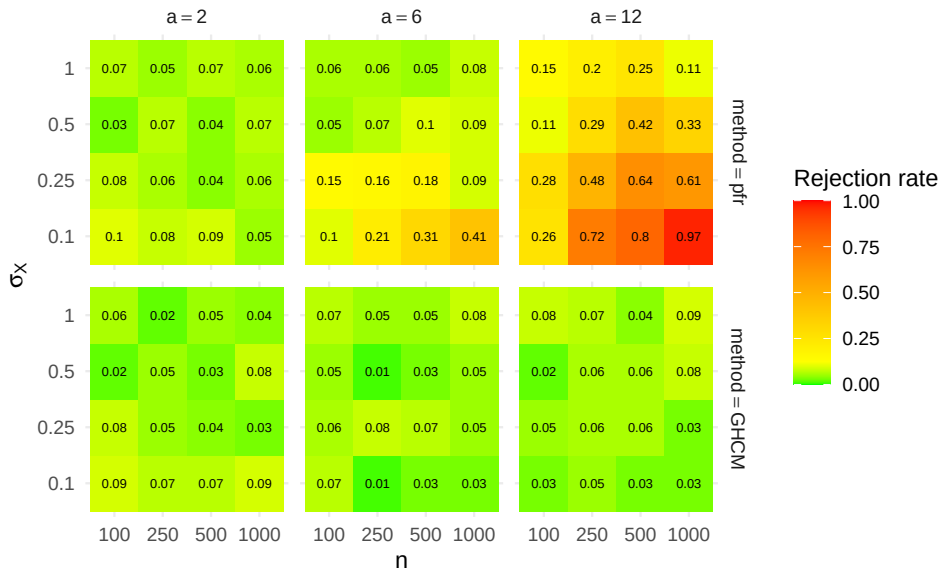


Fig. 2.1 Rejection rates in the various null settings considered in Section 2.5.1 for the nominal 5%-level `pfr` test (top) and GHCM test (bottom).

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 (2.22). We replace the regression model (2.23) for Y with

$$Y = \int_0^1 \alpha_a(t) Z(t)\,\mathrm{d}t + \int_0^1 \frac{\alpha_a(t)}{a} X(t)\,\mathrm{d}t + N_Y, \tag{2.25}$$

where $N_Y \sim \mathcal{N}_Y(0,1)$ 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 2.2. 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 2.1 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 2.10 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 2.3. 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.

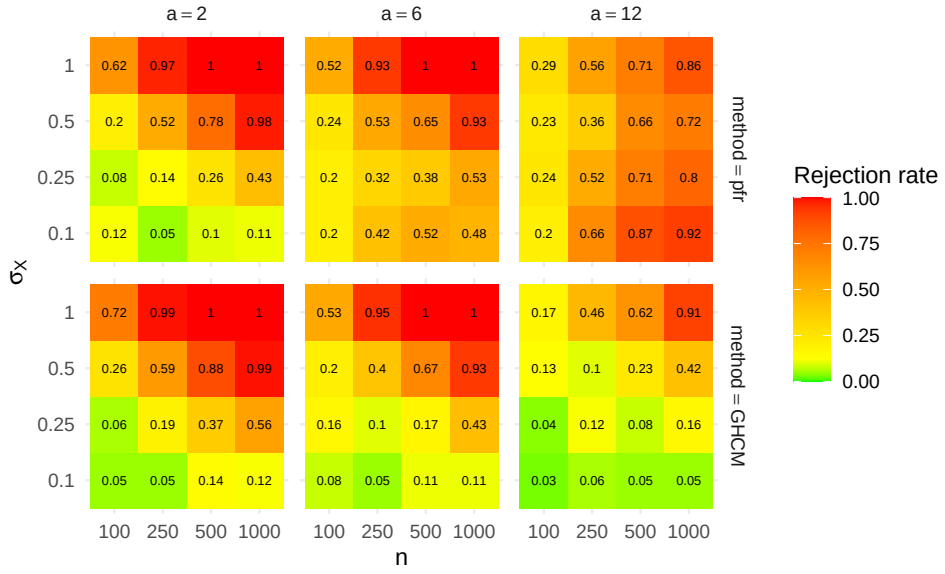


Fig. 2.2 Rejection rates in the various alternative settings considered in Section 2.5.1 (see (2.25)) for the nominal 5%-level `pfr` test (top) and GHCM test (bottom).

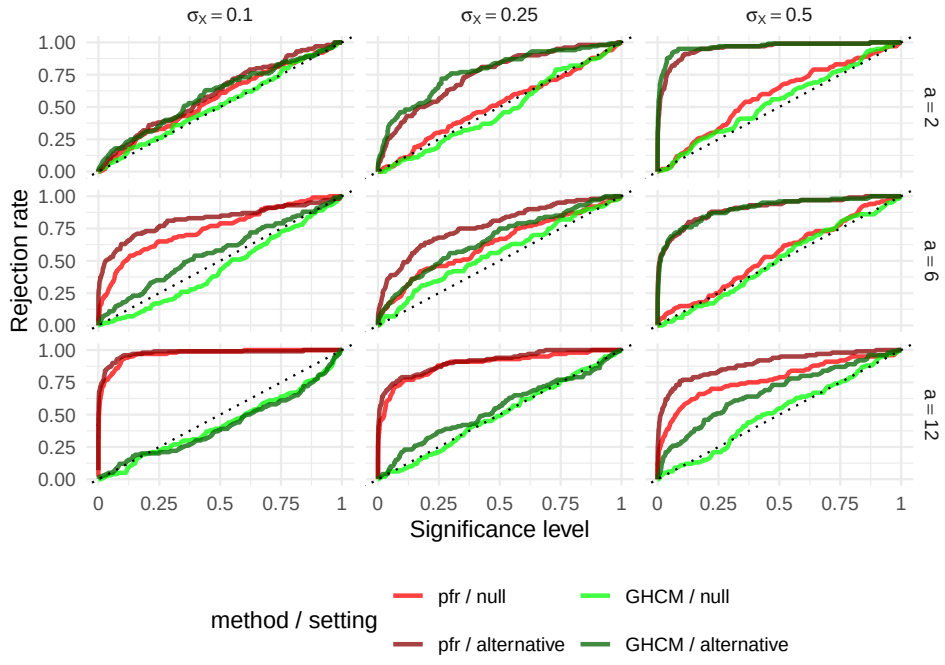


Fig. 2.3 Rejection rates against significance level for the `pfr` (red) and GHCM (green) tests under null (light) and alternative (dark) settings when $n = 500$.

In Section 2.10 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. The performance of the GHCM with `FDboost` is broadly similar to that displayed in Figures 2.1 and 2.2, 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 2.3, with the GHCM maintaining type I error control, and `pfr` tending to be anti-conservative in the more challenging settings.

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 2.5.1 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 regressing each of X and Y on Z will tend to be in similar directions making the equivalent of the term a_n in (2.2) 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 2.4. 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 2.10 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 2.3.2 and then apply the GHCM. We observe that the performance is similar to that in the fixed grid setting, though 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.

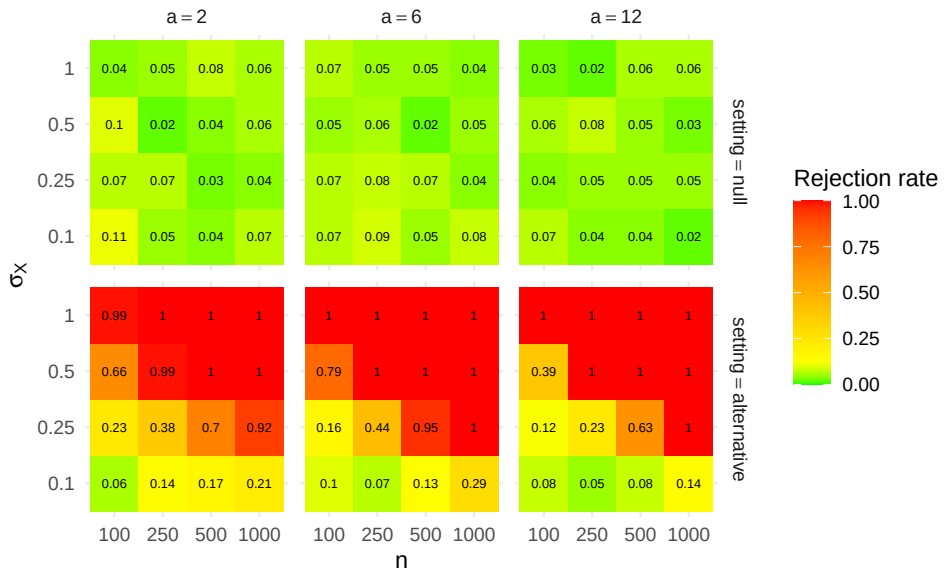


Fig. 2.4 Rejection rates in the various null (top) and alternative (bottom) settings considered in Section 2.5.1 for the nominal 5%-level GHCM test.

2.5.2 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 and Hooker, 2016)

$$Y = \int_0^\theta \alpha(t) X(t)\,\mathrm{d}t + \varepsilon, \tag{2.26}$$

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 (2.26) implies that defining the null hypotheses

$$H_{\tilde\theta}: \quad Y \perp\!\!\!\perp \{X(t)\}_{t>\tilde\theta} \,|\, \{X(t)\}_{t\le\tilde\theta} \tag{2.27}$$

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

$$\left[\inf\left\{\tilde\theta \in (0,1) \,:\, \psi \text{ accepts null } H_{\tilde\theta}\right\}, 1\right]. \tag{2.28}$$

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 (2.28) we initially consider the null hypothesis $H_{\tilde\theta}$ at 5 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 (2.26) with $\theta = 0.275, 0.675$ and with $\alpha(t) := 10(t+1)^{-1/3}$, X a

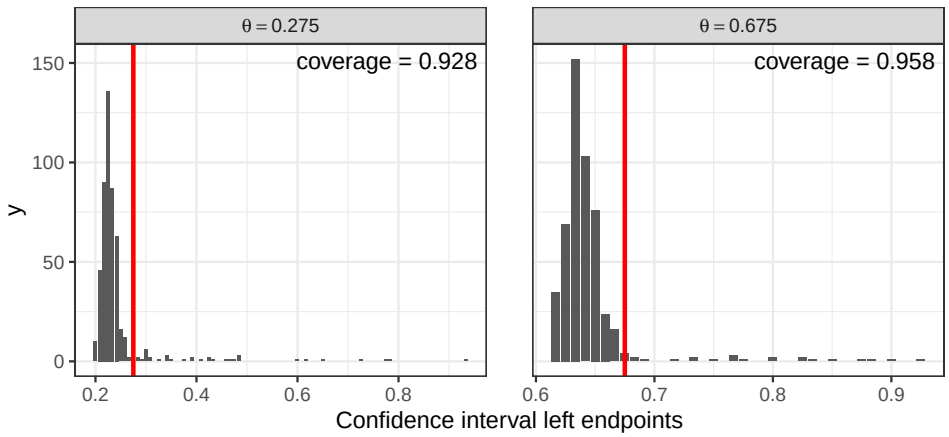


Fig. 2.5 Histograms of the left endpoints of 95% confidence intervals for truncation points $\theta = 0.275$ (left) and $\theta = 0.675$ (right), given by red vertical lines, in model (2.26) across 500 simulations.

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 2.5. We see that the empirical coverage probabilities are close to the nominal coverage of 95%.

2.5.3 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 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 2.3.

We illustrate this on an EEG dataset from a study on alcoholism (Ingber, 1997, 1998; Zhang et al., 1995). 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 $G_1, \ldots, G_{24}$ 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, $G_1 \cup \ldots \cup G_{24} = \{1, \ldots, 52\}$. We suppose the observed data are i.i.d. copies functional variables $(X_1, \ldots, X_{52})$, and then test the null hypothesis

$$X_{G_j} \perp\!\!\!\perp X_{G_k} \,|\, \{X_{G_m} : m \in \{1, \ldots, 24\} \setminus \{j, k\}\}, \tag{2.29}$$

for each $j, k \in \{1, \ldots, 24\}$ 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 $G_1, \ldots, G_{24}$, in which the edge between G_j and G_k, $j \neq k$ is missing if and only if (2.29) holds; that is, rejection of the null in (2.29) 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 (2.29) 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 and Silverman, 2005, Section 16) 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 2.10 of the supplement.

Figure 2.6 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. 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 and 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.

2.6 Conclusion

Testing the conditional independence $X \perp\!\!\!\perp Y \,|\, Z$ 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 and Peters, 2020). This hardness takes a more extreme form in the functional

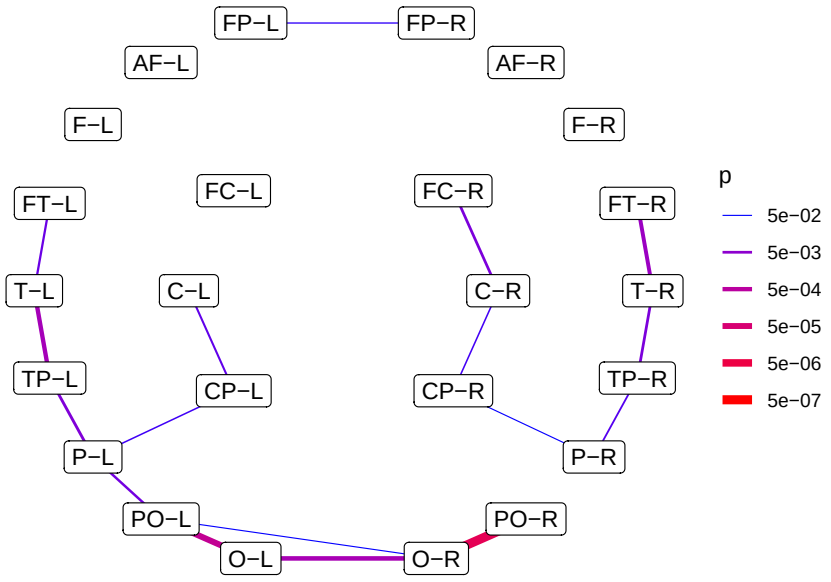


Fig. 2.6 Network summarising the output of conditional independence tests for each pair of groups. Only edges with p-values of less than 5% are shown with thicker lines indicating smaller p-values.

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.

2.7 Background on the hardness of functional Gaussian independence testing

In this section we provide the necessary background and prove the hardness result in Section 2.2. We use the notation and terminology described in the setup of Section 2.2 with the exception that $\mathcal{P}$, $\mathcal{P}_0$ and $\mathcal{Q}$ will consist of n i.i.d. copies of jointly Gaussian (X, Y, Z) rather than a single

copy. For a bounded linear operator $\mathscr{A}$ on a Hilbert space $\mathcal{H}$, we let $\mathscr{A}^*$ denote the adjoint of $\mathscr{A}$. For two orthogonal subspaces $\mathcal{A}$ and $\mathcal{B}$ of a Hilbert space $\mathcal{H}$, we write $\mathcal{A} \oplus \mathcal{B}$ for the orthogonal direct sum of $\mathcal{A}$ and $\mathcal{B}$.

In Section 2.7.1 we consider the setup of Section 2.2 in the specific case where all the Hilbert spaces are finite-dimensional. We show that for any $Q \in \mathcal{Q}$, sample size n and $\varepsilon > 0$, we can find a sufficiently large dimension of $\mathcal{H}_Z$ such that any test of size α over $\mathcal{P}_0^Q$ has power at most $\alpha + \varepsilon$ against any alternative. In Section 2.7.2 we use this to prove Theorem 2.1. In Section 2.7.3 we review the theory of regular conditional probabilities and conditional distributions of Hilbertian random variables and prove several Hilbertian analogues of well-known multivariate Gaussian results. Sections 2.7.1 and 2.7.2 except Lemma 2.1 contain new material while Section 2.7.3 is primarily a review of relatively well-known results.

2.7.1 Power of finite-dimensional Gaussian conditional independence testing

Before we consider Gaussian conditional independence testing, we present the following general result from Kraft (1955). A summary is given in LeCam (1973).

Lemma 2.1. *Let $\mathcal{P}$ and $\mathcal{Q}$ denote two families of probability measures on some measurable space $(\mathcal{X}, \mathcal{A})$ and assume that both families are dominated by a σ-finite measure. Consider the problem of testing the null hypothesis that the given data is from a distribution in $\mathcal{P}$ against the alternative that the distribution is in $\mathcal{Q}$. Let d_{TV} denote the total variation distance and $\widetilde{\mathcal{P}}$ and $\widetilde{\mathcal{Q}}$ the closed convex hulls of $\mathcal{P}$ and $\mathcal{Q}$. Then*

$$\inf_{\psi:\mathcal{X}\to[0,1]} \sup_{P\in\mathcal{P},Q\in\mathcal{Q}} \left[\int \psi \,\mathrm{d}P + \int (1-\psi)\,\mathrm{d}Q\right] = 1 - \inf_{P\in\widetilde{\mathcal{P}},Q\in\widetilde{\mathcal{Q}}} d_{TV}(P,Q).$$

An immediate consequence of this is that for any test ψ that has size α and power function $\beta : \mathcal{Q} \to [0,1]$, $\beta(Q) = \int \psi \,\mathrm{d}Q$, we have

$$\inf_{Q\in\mathcal{Q}} \beta(Q) \leq \alpha + \inf_{P\in\widetilde{\mathcal{P}},Q\in\widetilde{\mathcal{Q}}} d_{TV}(P,Q) \leq \alpha + \inf_{P\in\widetilde{\mathcal{P}},Q\in\mathcal{Q}} d_{TV}(P,Q).$$

In most practical situations both $\mathcal{P}$ and $\mathcal{Q}$ will consist of product measures on a product space corresponding to a situation where we observe a sample of n i.i.d. observations of some random variable. The theorem states that a lower bound on the sum of the type I and type II error probabilities of testing the null that data is from a distribution in $\mathcal{P}$ against the alternative that the distribution is in $\mathcal{Q}$ is given by 1 minus the total variation distance between the closed convex hulls of $\mathcal{P}$ and $\mathcal{Q}$. As a consequence we see that the power of a test is upper bounded by the size plus the total variation distance between the closed convex hull of $\mathcal{P}$ and $\mathcal{Q}$.

In the remainder of this section we will consider the testing problem described in Section 2.2 with $\mathcal{H}_X = \mathbb{R}^{d_X}$ and $\mathcal{H}_Z = \mathbb{R}^{d_Z}$ for $d_X, d_Z \in \mathbb{N}$. To produce bounds on the power of a test in this setting, we will construct an explicit TV-approximation to a family of particularly simple distributions in $\mathcal{Q}$ using a distribution in the convex hull of the null distributions. We will need the following upper bound on the total variation distance between measures.

Lemma 2.2. *Let P and Q be probability measures where P has density f with respect to Q. Then*

$$4d_{\mathrm{TV}}(P,Q)^2 \leq \int f^2 \,\mathrm{d}Q - 1.$$

Proof. We may assume that the integral of f^2 with respect to Q is finite, otherwise the inequality is trivially valid. Then by Jensen's inequality, we get

$$d_{\mathrm{TV}}(P,Q)^2 = \frac{1}{4}\left(\int |f-1| \,\mathrm{d}Q\right)^2 \leq \frac{1}{4}\int (f-1)^2 \,\mathrm{d}Q = \frac{1}{4}\int f^2 \,\mathrm{d}Q - \frac{1}{4}. \qquad \square$$

Using this bound and Lemma 2.1, we can show the following result.

Theorem 2.6. *Let Q be a distribution consisting of n i.i.d. copies of jointly Gaussian (X,Y,Z) on $(\mathbb{R},\mathbb{R},\mathbb{R}^d)$ for some $d \in \mathbb{N}$, where X and Y are standard Gaussian, Z is mean zero with identity covariance matrix, $\mathrm{Cov}(X,Z) = \mathrm{Cov}(Y,Z) = 0$ and $\mathrm{Cov}(X,Y) = \rho \in (0,1)$. Consider the testing problem described in Section 2.2 with $\mathcal{H}_X = \mathbb{R}$ and $\mathcal{H}_Z = \mathbb{R}^d$ and let ψ be the test function of a size α test over $\mathcal{P}_0^Q$. Writing β for the power of ψ against Q, we have*

$$\beta \leq \alpha + \frac{1}{2}\sqrt{-1 + (1+\rho)^n \sum_{k=0}^{d} \frac{\binom{d}{k}}{2^d(1+(3-4k/d)\rho)^n}}.$$

In particular, for fixed n the upper bound converges to α as d increases.

Proof. Let $\tau \in \{-1,1\}^d$ and let P_τ denote the Gaussian distribution consisting of n i.i.d. copies of jointly Gaussian (X,Y,Z) where X and Y are standard Gaussian, Z is mean zero with identity covariance matrix, $\mathrm{Cov}(X,Y) = \rho$ and $\mathrm{Cov}(X,Z) = \mathrm{Cov}(Y,Z) = \sqrt{\frac{\rho}{d}}\tau^\top$. For every $\tau \in \{-1,1\}^d$, it is clear that $X \perp\!\!\!\perp Y \,|\, Z$ under P_τ and thus forming

$$P := \frac{1}{2^d}\sum_{\tau\in\{-1,1\}^d} P_\tau$$

we note that P is in the closed convex hull of the set of null distributions. Let Γ_τ and Γ_Q denote the $n(d+2)$-dimensional covariance matrices of the n i.i.d. copies of (X,Y,Z) under P_τ and Q respectively. These are block-diagonal, and we let Σ_τ and Σ_Q respectively denote the matrices in the diagonal, corresponding to the covariance of a single observation of (X,Y,Z) under P_τ and Q. By standard manipulations of densities, the density of P with respect to Q is simply the ratio of their respective densities with respect to the Lebesgue measure. We have

$$\Sigma_\tau = \begin{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} & \sqrt{\frac{\rho}{d}}\begin{pmatrix} \tau^\top \\ \tau^\top \end{pmatrix} \\ \sqrt{\frac{\rho}{d}}\begin{pmatrix} \tau & \tau \end{pmatrix} & I_d \end{pmatrix}$$

and, letting I_d denote the d-dimensional identity matrix,

$$\Sigma_Q = \begin{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} & 0 \\ 0 & I_d \end{pmatrix}.$$

The determinant of Σ_Q is $1-\rho^2$ by Laplace-expanding the first row. Letting J_2 denote the 2-dimensional matrix of ones, we have

$$\det(\Sigma_\tau) = \det(I_d)\det\left(\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \rho J_2\right) = (1-\rho)^2$$

by Schur's formula. Defining f to be the density of P with respect to Q, we see that

$$f(v) = \frac{1}{2^d}\frac{(1+\rho)^{n/2}}{(1-\rho)^{n/2}} \sum_{\tau\in\{-1,1\}^d} \exp\left(-\frac{1}{2}v^\top(\Gamma_\tau^{-1} - \Gamma_Q^{-1})v\right)$$

since the determinants of Γ_τ and Γ_Q are the determinants of Σ_τ and Σ_Q to the nth power. From this we get that

$$\int f^2 \,\mathrm{d}Q = \frac{1}{2^{2d}}\frac{(1+\rho)^n}{(1-\rho)^n} \sum_{\tau,\tau'\in\{-1,1\}^d} \int \exp\left(-\frac{1}{2}v^\top(\Gamma_\tau^{-1} + \Gamma_{\tau'}^{-1} - 2\Gamma_Q^{-1})v\right) \mathrm{d}Q(v) =$$
$$\frac{1}{2^{2d}}\frac{(1+\rho)^n}{(1-\rho)^n}\frac{1}{\sqrt{(2\pi)^{n(d+2)}(1-\rho^2)^n}} \sum_{\tau,\tau'\in\{-1,1\}^d} \int \exp\left(-\frac{1}{2}v^\top(\Gamma_\tau^{-1} + \Gamma_{\tau'}^{-1} - \Gamma_Q^{-1})v\right) \mathrm{d}\lambda_{n(d+2)}(v),$$

where $\lambda_{n(d+2)}$ denotes the $n(d+2)$-dimensional Lebesgue measure. Each integral is the integral of an unnormalised Gaussian density in $\mathbb{R}^{n(d+2)}$, and thus we can simplify further to get

$$\begin{aligned}
\int f^2 \,\mathrm{d}Q &= \frac{1}{2^{2d}}\frac{(1+\rho)^n}{(1-\rho)^n}\frac{1}{(1-\rho^2)^{n/2}} \sum_{\tau,\tau'\in\{-1,1\}^d} \sqrt{\det\left[(\Gamma_\tau^{-1} + \Gamma_{\tau'}^{-1} - \Gamma_Q^{-1})^{-1}\right]} \\
&= \frac{1}{2^{2d}}\frac{(1+\rho)^n}{(1-\rho)^n}\frac{1}{(1-\rho^2)^{n/2}} \sum_{\tau,\tau'\in\{-1,1\}^d} \det(\Gamma_\tau^{-1} + \Gamma_{\tau'}^{-1} - \Gamma_Q^{-1})^{-1/2} \\
&= \frac{1}{2^{2d}}\frac{(1+\rho)^n}{(1-\rho)^n}\frac{1}{(1-\rho^2)^{n/2}} \sum_{\tau,\tau'\in\{-1,1\}^d} \det(\Sigma_\tau^{-1} + \Sigma_{\tau'}^{-1} - \Sigma_Q^{-1})^{-n/2},
\end{aligned}$$

by again using the block diagonal structure of Γ_Q and the Γ_τ's. Recall that for a symmetric block matrix

$$\begin{pmatrix} A & B^\top \\ B & C \end{pmatrix}^{-1} = \begin{pmatrix} (A - B^\top C^{-1}B)^{-1} & -(A - B^\top C^{-1}B)^{-1}B^\top C^{-1} \\ -C^{-1}B(A - B^\top C^{-1}B^\top)^{-1} & C^{-1} + C^{-1}B(A - B^\top C^{-1}B)^{-1}B^\top C^{-1} \end{pmatrix}.$$

Using this, we see that

$$\Sigma_Q^{-1} = \begin{pmatrix} \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} & 0 \\ 0 & I_d \end{pmatrix}$$

and

$$\Sigma_\tau^{-1} = \begin{pmatrix} \frac{1}{1-\rho} I_2 & -\frac{1}{1-\rho}\sqrt{\frac{\rho}{d}}\begin{pmatrix} \tau^\top \\ \tau^\top \end{pmatrix} \\ -\frac{1}{1-\rho}\sqrt{\frac{\rho}{d}}\begin{pmatrix} \tau & \tau \end{pmatrix} & I_d + \frac{2\rho}{(1-\rho)d}\tau\tau^\top \end{pmatrix}.$$

Further,

$$\Sigma_\tau^{-1} + \Sigma_{\tau'}^{-1} - \Sigma_Q^{-1} = \begin{pmatrix} A & B^\top \\ B & C \end{pmatrix},$$

where

$$\begin{aligned} A &:= \frac{1}{1-\rho^2}\begin{pmatrix} 2\rho+1 & \rho \\ \rho & 2\rho+1 \end{pmatrix} \\ B &:= -\frac{1}{1-\rho}\sqrt{\frac{\rho}{d}}\begin{pmatrix} \tau+\tau' & \tau+\tau' \end{pmatrix} \\ C &:= I_d + \frac{2\rho}{(1-\rho)d}(\tau\tau^\top + \tau'\tau'^\top). \end{aligned}$$

We may once more use Schur's formula for the determinant of a block matrix to find that

$$\det(\Sigma_\tau^{-1} + \Sigma_{\tau'}^{-1} - \Sigma_Q^{-1}) = \det(C)\det(A - B^\top C^{-1} B).$$

Defining $V = \begin{pmatrix} \tau & \tau' \end{pmatrix}$, we note that $C = I_d + \frac{2\rho}{(1-\rho)d} VV^\top$ and defining further

$$M := I_2 + \frac{2\rho}{(1-\rho)d} V^\top V = \frac{1}{d(1-\rho)}\begin{pmatrix} d(1+\rho) & 2\rho\langle\tau,\tau'\rangle \\ 2\rho\langle\tau,\tau'\rangle & d(1+\rho), \end{pmatrix}$$

the Weinstein–Aronszajn identity yields that

$$\det(C) = \det(M) = \frac{(d(1+\rho) + 2\rho\langle\tau,\tau'\rangle)(d(1+\rho) - 2\rho\langle\tau,\tau'\rangle)}{d^2(1-\rho)^2}.$$

The Woodbury matrix identity yields that

$$C^{-1} = I_d - \frac{2\rho}{(1-\rho)d} V M^{-1} V^\top.$$

Hence,

$$\det(A - B^\top C^{-1} B) = \det\left(A - B^\top B + \frac{2\rho}{(1-\rho)d} B^\top V M^{-1} V^\top B\right).$$

Now

$$M^{-1} = \frac{(1-\rho)d}{(d(1+\rho) + 2\rho\langle\tau,\tau'\rangle)(d(1+\rho) - 2\rho\langle\tau,\tau'\rangle)}\begin{pmatrix} d(1+\rho) & -2\rho\langle\tau,\tau'\rangle \\ -2\rho\langle\tau,\tau'\rangle & d(1+\rho) \end{pmatrix}$$

and

$$B^\top V = -\frac{1}{1-\rho}\sqrt{\frac{\rho}{d}}(d + \langle \tau, \tau' \rangle) J_2,$$

where J_2 is the 2-dimensional matrix of ones. Thus,

$$\frac{2\rho}{(1-\rho)d} B^\top V M^{-1} V^\top B = \frac{2\rho^2(d+\langle \tau,\tau'\rangle)^2}{(1-\rho)^3 d^2} J_2 M^{-1} J_2 = \frac{4\rho^2(d+\langle \tau,\tau'\rangle)^2}{(1-\rho)^2 d(d(1+\rho)+2\rho\langle \tau,\tau'\rangle)} J_2.$$

Since

$$B^\top B = \frac{2\rho}{(1-\rho)^2 d}(d + \langle \tau, \tau' \rangle) J_2$$

we get that

$$\begin{aligned}
&\det(A - B^\top C^{-1} B) \\
&= \det\left(A + \left(\frac{4\rho^2(d+\langle \tau,\tau'\rangle)^2}{(1-\rho)^2 d(d(1+\rho)+2\rho\langle \tau,\tau'\rangle)} - \frac{2\rho}{(1-\rho)^2 d}(d+\langle \tau,\tau'\rangle) \right) J_2 \right) \\
&= \det\left(A - \frac{2\rho(d+\langle \tau,\tau'\rangle)}{(1-\rho)(d(1+\rho)+2\rho\langle \tau,\tau'\rangle)} J_2 \right) \\
&= \frac{\det\left((d(1+\rho)+2\rho\langle \tau,\tau'\rangle) \begin{pmatrix} 2\rho+1 & \rho \\ \rho & 2\rho+1 \end{pmatrix} - 2\rho(d+\langle \tau,\tau'\rangle)(1+\rho) J_2 \right)}{(1-\rho)^2(1+\rho)^2(d(1+\rho)+2\rho\langle \tau,\tau'\rangle)^2} \\
&= \frac{\det\begin{pmatrix} (d(1+\rho)+2\rho\langle \tau,\tau'\rangle)\rho + (1+\rho)(1-\rho)d & (1+\rho)(1-\rho)d - (d(1+\rho)+2\rho\langle \tau,\tau'\rangle) \\ (1+\rho)(1-\rho)d - (d(1+\rho)+2\rho\langle \tau,\tau'\rangle) & (d(1+\rho)+2\rho\langle \tau,\tau'\rangle)\rho + (1+\rho)(1-\rho)d \end{pmatrix}}{(1-\rho)^2(1+\rho)^2(d(1+\rho)+2\rho\langle \tau,\tau'\rangle)^2} \\
&= \frac{(d(1+\rho)+2\rho\langle \tau,\tau'\rangle)(1+\rho)(\rho-1) + 2(1+\rho)^2(1-\rho)d}{(1-\rho)^2(1+\rho)^2(d(1+\rho)+2\rho\langle \tau,\tau'\rangle)} \\
&= \frac{d(1+\rho) - 2\rho\langle \tau,\tau'\rangle}{(1-\rho)(1+\rho)(d(1+\rho)+2\rho\langle \tau,\tau'\rangle)}
\end{aligned}$$

and thus

$$\det(\Sigma_\tau^{-1} + \Sigma_{\tau'}^{-1} - \Sigma_Q^{-1}) = \frac{(d(1+\rho) - 2\rho\langle \tau,\tau'\rangle)^2}{d^2(1-\rho)^3(1+\rho)}.$$

Returning to the squared integral of f^2 with respect to Q, we get that

$$\begin{aligned}
\int f^2 \,\mathrm{d}Q &= \frac{1}{2^{2d}} \frac{(1+\rho)^n}{(1-\rho)^n} \frac{1}{(1-\rho^2)^{n/2}} \sum_{\tau,\tau' \in \{-1,1\}^d} \frac{d^n \sqrt{(1-\rho)^{3n}(1+\rho)^n}}{|d(1+\rho) - 2\rho\langle \tau,\tau'\rangle|^n} \\
&= \frac{1}{2^{2d}}(1+\rho)^n \sum_{\tau,\tau' \in \{-1,1\}^d} \frac{d^n}{|d(1+\rho) - 2\rho\langle \tau,\tau'\rangle|^n}.
\end{aligned}$$

For $\tau, \tau' \in \{-1,1\}^d$, $\langle \tau, \tau' \rangle = 2k - d$ where k is the number of indices where $\tau_i = \tau'_i$. Thus instead of summing over $\tau, \tau' \in \{-1,1\}^d$, we can count the number of (τ, τ')-pairs where τ and τ' agree in exactly k positions. For each τ, there are $\binom{d}{k}$ other elements in $\{-1,1\}^d$ agreeing in

exactly k positions and there are 2^d different τ's, hence

$$\int f^2 \,\mathrm{d}Q = \frac{1}{2^{2d}}(1+\rho)^n \sum_{k=0}^{d} \frac{d^n \binom{d}{k} 2^d}{|d(1+\rho) - 2\rho(2k-d)|^n}$$
$$= (1+\rho)^n \sum_{k=0}^{d} \frac{d^n \binom{d}{k}}{2^d (d+\rho(3d-4k))^n} = (1+\rho)^n \sum_{k=0}^{d} \frac{\binom{d}{k}}{2^d (1+\rho(3-4k/d))^n}.$$

The result now follows from Proposition 2.2 and Lemma 2.1.

To see this for each n the bound converges to α as d increases, let W_d be a random variable with a binomial distribution with probability parameter $1/2$ and with d trials and note that

$$\sum_{k=0}^{d} \frac{\binom{d}{k}}{2^d (1+\rho(3-4k/d))^n} = \mathbb{E}\left((1+\rho(3-4W_d/d))^{-n}\right).$$

By the Strong Law of Large Numbers (SLLN), $W_d/d \xrightarrow{a.s.} 1/2$ and thus $(1+\rho(3-4W_d/d))^{-n} \xrightarrow{a.s.} (1+\rho)^{-n}$. Since $(1+\rho(3-4W_d/d))^{-n} \leq (1-\rho)^{-n}$, we get by the bounded convergence theorem that

$$\lim_{d\to\infty} \mathbb{E}\left((1+\rho(3-4W_d/d))^{-n}\right) = \mathbb{E}\left((1+\rho)^{-n}\right) = (1+\rho)^{-n},$$

and hence the upper bound on the power converges to α. □

We can generalise the previous result to the situation where X and Y are of arbitrary finite dimension.

Theorem 2.7. *Let Q be a distribution consisting of n i.i.d. copies of jointly Gaussian (X, Y, Z) on $(\mathbb{R}^{d_X}, \mathbb{R}^{d_Y}, \mathbb{R}^{d_Z})$ for some $d_X, d_Y, d_Z \in \mathbb{N}$ where X, Y and Z are all mean zero with identity covariance matrix, $\mathrm{Cov}(X,Z) = \mathrm{Cov}(Y,Z) = 0$ and $\mathrm{Cov}(X,Y) = R$ for some rectangular diagonal matrix R with diagonal entries $\rho_1, \ldots, \rho_r \in (0,1)$, where $r = \min(d_X, d_Y)$. Consider the testing problem described in Section 2.2 with $\mathcal{H}_X = \mathbb{R}^{d_X}$ and $\mathcal{H}_Z = \mathbb{R}^{d_Z}$ and let ψ be the test function of a size α test over $\mathcal{P}_0^Q$. Assume that $d_Z \geq r$ and let $d = \lfloor d_Z/r \rfloor$. Letting β denote the power of ψ against Q, we have*

$$\beta \leq \alpha + \frac{1}{2}\sqrt{-1 + \prod_{i=1}^{r}\left((1+\rho_i)^n \sum_{k=0}^{d} \frac{\binom{n}{k}}{2^d (1+(3-4k/d)\rho_i)^n}\right)}.$$

In particular for fixed n the upper bound converges to α as d_Z increases.

Proof. Assume without loss of generality that $d_X \geq d_Y$. The proof follows a similar idea to the proof of Theorem 2.6. In what follows we consider a different ordering of the variables than the natural one given by (X, Y, Z). We consider $r+1$ blocks, where the first r blocks are $(X_i, Y_i, Z_{(i-1)d+1}, \ldots, Z_{id})$ for $i \in \{1, \ldots, r\}$ and the final block consists of the remaining components of X and Z. When we consider n i.i.d. copies, we will again reorder the variables such that we consider each block separately. As a consequence of doing this, the covariance matrix of n i.i.d. copies under Q, Ξ_Q, can be written as a block-diagonal matrix with r

$n(d+2) \times n(d+2)$ blocks $\Gamma_{Q,i}$ and a final identity matrix block. Each of the $\Gamma_{Q,i}$'s is again a block-diagonal matrix consisting of n identical blocks $\Sigma_{Q,i}$ of the form

$$\Sigma_{Q,i} = \begin{pmatrix} \begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix} & 0 \\ 0 & I_d \end{pmatrix}.$$

Let now $\mathcal{T} = (\{-1,1\}^d)^r$ and for each $\tau = (\tau_1, \ldots, \tau_r) \in \mathcal{T}$ let P_τ denote the Gaussian distribution consisting of n i.i.d. copies of jointly Gaussian (X, Y, Z) where X, Y and Z are mean zero with identity covariance, $\mathrm{Cov}(X,Y) = R$ and $\mathrm{Cov}(X,Z) = \mathrm{Cov}(Y,Z) = 0$ except for

$$\mathrm{Cov}(X_i, (Z_{(i-1)d+1}, \ldots, Z_{id})) = \mathrm{Cov}(Y_i, (Z_{(i-1)d+1}, \ldots, Z_{id})) = \sqrt{\frac{\rho_i}{d}}\tau_i^\top$$

for $i \in \{1, \ldots, r\}$. Arranging the random variables as before, the covariance matrix of n i.i.d. copies under P_τ, Ξ_τ, is a block-diagonal matrix with r $n(d+2) \times n(d+2)$ blocks $\Gamma_{\tau,i}$ and a final identity matrix block. Each of the $\Gamma_{\tau,i}$'s is again a block-diagonal matrix consisting of n identical blocks $\Sigma_{\tau,i}$ of the form

$$\Sigma_{\tau,i} = \begin{pmatrix} \begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix} & \sqrt{\frac{\rho_i}{d}}\tau_i^\top \\ \sqrt{\frac{\rho_i}{d}}\tau_i & I_d \end{pmatrix}.$$

Clearly $X \perp\!\!\!\perp Y \,|\, Z$ under P_τ for every $\tau \in \mathcal{T}$ and thus letting

$$P := \frac{1}{2^{dr}} \sum_{\tau \in \mathcal{T}} P_\tau,$$

we note that P is in the closed convex hull of the null distributions. Letting f be the density of P with respect to Q, we see that

$$f(v) = \frac{1}{2^{dr}} \left(\prod_{i=1}^r \frac{1+\rho_i}{1-\rho_i} \right)^{n/2} \sum_{\tau \in \mathcal{T}} \exp\left(-\frac{1}{2} v^\top (\Xi_\tau^{-1} - \Xi_Q^{-1}) v \right),$$

since this is simply the ratio of their respective densities with respect to the Lebesgue measure. We can now repeat the argument of the proof of Theorem 2.6 to obtain

$$\int f^2 \, \mathrm{d}Q = \frac{1}{2^{2dr}} \prod_{i=1}^r \left(\frac{1+\rho_i}{(1-\rho_i)\sqrt{1-\rho_i^2}} \right)^n \sum_{\tau, \tau' \in \mathcal{T}} \left(\sqrt{\det(\Xi_\tau^{-1} + \Xi_{\tau'}^{-1} - \Xi_Q^{-1})} \right)^{-1}.$$

The determinant can be written as

$$\det(\Xi_\tau^{-1} + \Xi_{\tau'}^{-1} - \Xi_Q^{-1}) = \prod_{i=1}^r \det(\Gamma_{\tau,i}^{-1} + \Gamma_{\tau',i}^{-1} - \Gamma_{Q,i}^{-1})$$

by the block-diagonal structure of the Ξ's. In the proof of Theorem 2.6, we derive that

$$\det(\Gamma_{\tau,i}^{-1} + \Gamma_{\tau',i}^{-1} - \Gamma_{Q,i}^{-1}) = \left(\frac{(d(1+\rho_i) - 2\rho_i\langle\tau_i, \tau_i'\rangle)^2}{d^2(1+\rho_i)(1-\rho_i)^3}\right)^n.$$

Therefore,

$$\int f^2 \, \mathrm{d}Q = \frac{1}{2^{2dr}} \left(\prod_{j=1}^{r}(1+\rho_j)^n\right) \sum_{\tau,\tau' \in \mathcal{T}} \prod_{i=1}^{r} \frac{d^n}{|d(1+\rho_i) - 2\rho_i\langle\tau_i, \tau_i'\rangle|^n}.$$

Since each factor of the second product only depends on the ith component of τ and τ', we can interchange the product and sum and apply the same counting arguments as in Theorem 2.6 to get that

$$\int f^2 \, \mathrm{d}Q = \prod_{i=1}^{r}\left((1+\rho_i)^n \sum_{k=0}^{d} \frac{\binom{n}{k}}{2^d(1+(3-4k/d)\rho_i)^n}\right)$$

as desired. We can repeat the same SLLN-based limiting arguments as in Theorem 2.6 to show that as d increases the integral will converge to 1 and hence the power is bounded by the size in the limit. □

Having shown that for each n and d, we have an upper bound on the power of a Gaussian conditional independence test against a simple alternative, we can now show this also holds for Gaussian conditional independence testing problems against other Q.

Lemma 2.3. *Let $Q \in \mathcal{Q}$ be a distribution consisting of n i.i.d. copies of jointly Gaussian and injective (X, Y, Z) on $(\mathbb{R}^{d_X}, \mathbb{R}^{d_Y}, \mathbb{R}^{d_Z})$ with non-singular covariance for some $d_X, d_Y, d_Z \in \mathbb{N}$. Consider the testing problem described in Section 2.2 with $\mathcal{H}_X = \mathbb{R}^{d_X}$ and $\mathcal{H}_Z = \mathbb{R}^{d_Z}$ and let ψ be the test function of a size α test over $\mathcal{P}_0^Q$ with power β against Q. Then there exists a $d_X \times d_Y$-rectangular diagonal matrix R with diagonal entries $\rho_1, \ldots, \rho_r \in (0,1)$, a distribution $\tilde{Q}$ consisting of n i.i.d. copies of jointly Gaussian $(\tilde{X}, \tilde{Y}, \tilde{Z})$ where $\tilde{X}$, $\tilde{Y}$ and $\tilde{Z}$ are all mean zero with identity covariance matrix, $\mathrm{Cov}(\tilde{X}, \tilde{Z}) = \mathrm{Cov}(\tilde{Y}, \tilde{Z}) = 0$ and $\mathrm{Cov}(\tilde{X}, \tilde{Y}) = R$ and a test size α test over $\mathcal{P}_0^{\tilde{Q}}$ with power β against $\tilde{Q}$.*

Proof. Let ψ denote the test function of the test with power β against Q and μ and Σ denote the mean and covariance matrix of (X, Y, Z) under Q. We construct a new test with test function $\tilde{\psi}$ performed by first applying a transformation f to each sample of the data and then applying ψ. The transformation $f : \mathbb{R}^{d_X+d_Y+d_Z} \to \mathbb{R}^{d_X+d_Y+d_Z}$ is an affine transformation given by $f(v) = Ax + \mu$ where

$$A = \begin{pmatrix} D & M \\ 0 & B \end{pmatrix}$$

for a block-diagonal matrix D consisting of a $d_X \times d_X$ matrix D_X and $d_Y \times d_Y$ matrix D_Y, a $(d_X + d_Y) \times d_Z$ matrix M and a full rank $d_Z \times d_Z$ matrix B.

Note first that such a transformation preserves conditional independence. Let (X^0, Y^0, Z^0) be jointly Gaussian with $X^0 \perp\!\!\!\perp Y^0 \,|\, Z^0$, joint mean μ^0 and covariance matrix Σ^0. The distribution of $(\check{X}_0, \check{Y}_0, \check{Z}_0) := f(X^0, Y^0, Z^0)$ is again Gaussian by the finite-dimensional version of

Proposition 2.6 and has mean $A\mu^0 + \mu$ and covariance

$$A\Sigma^0 A^\top = \begin{pmatrix} D\Sigma^0_{XY}D^\top + M\Sigma^0_{Z,XY}D^\top + D\Sigma^0_{XY,Z}M^\top + M\Sigma^0_Z M^\top & D\Sigma^0_{XY,Z}A^\top + M\Sigma^0_Z B^\top \\ B\Sigma^0_{Z,XY}D^\top + B\Sigma^0_Z M^\top & B\Sigma^0_Z B^\top \end{pmatrix},$$

where $\Sigma^0_{XY} = \mathrm{Cov}((X^0, Y^0))$, $\Sigma^0_{XY,Z} = \Sigma^0_{Z,XY} = \mathrm{Cov}((X^0, Y^0), Z^0)$ and $\Sigma^0_Z = \mathrm{Cov}(Z_0)$. Using the finite-dimensional version of Proposition 2.7, we get that the conditional distribution of $(\check{X}_0, \check{Y}_0)$ given $\check{Z}_0$ is again Gaussian with covariance matrix

$$\begin{aligned} &D\Sigma^0_{XY}D^\top + M\Sigma^0_{Z,XY}D^\top + D\Sigma^0_{XY,Z}M^\top + M\Sigma^0_Z M^\top \\ &\quad - (D\Sigma^0_{XY,Z}B^\top + M\Sigma^0_Z B^\top)(B\Sigma^0_Z B^\top)^{-1}(B\Sigma^0_{Z,XY}D^\top + B\Sigma^0_Z M^\top) \\ =&D(\Sigma^0_{XY} - \Sigma^0_{XY,Z}\Sigma^0_Z\Sigma^0_{Z,XY})D^\top. \end{aligned}$$

The matrix $\Sigma^0_{XY} - \Sigma^0_{XY,Z}\Sigma^0_Z\Sigma^0_{Z,XY}$ is the conditional covariance matrix of (X^0, Y^0) given Z^0 and is block-diagonal since $X^0 \perp\!\!\!\perp Y^0 \,|\, Z^0$ by the multivariate analogue of Proposition 2.5. By the same proposition, since D is block-diagonal, we see that the conditional covariance of $(\check{X}_0, \check{Y}_0)$ given $\check{Z}_0$ is block-diagonal and hence $\check{X}_0 \perp\!\!\!\perp \check{Y}_0 \,|\, \check{Z}_0$ as desired.

Let now

$$\Sigma^{-1/2}_{X|Z}\Sigma_{XY\,|\,Z}\Sigma^{-1/2}_{Y\,|\,Z} = USV^\top$$

be the singular-value decomposition of the normalised conditional covariance of X and Y given Z under Q. The normalisation ensures that S is a rectangular diagonal matrix with diagonal entries in the open unit interval. Let

$$B := \Sigma^{1/2}_Z, \quad M := \begin{pmatrix} \Sigma_{X,Z}\Sigma^{-1/2}_Z \\ \Sigma_{Y,Z}\Sigma^{-1/2}_Z \end{pmatrix}$$
$$D := \begin{pmatrix} \Sigma^{1/2}_{X\,|\,Z}U & 0 \\ 0 & \Sigma^{1/2}_{Y\,|\,Z}V \end{pmatrix}, \quad R := S$$

and $(\check{X}, \check{Y}, \check{Z}) = f((\tilde{X}, \tilde{Y}, \tilde{Z}))$ where $(\tilde{X}, \tilde{Y}, \tilde{Z}) \sim \tilde{Q}$. Proposition 2.6 yields that $(\check{X}, \check{Y}, \check{Z}) \sim Q$ and hence when applying ψ, we have power β by assumption. Since A also transforms a null distribution with identity covariance into a null distribution with where Z has mean μ_Z and covariance Σ_Z, we have the desired result. □

2.7.2 Hardness of infinite-dimensional Hilbertian Gaussian conditional independence testing

In this section we consider the testing problem described in Section 2.2 with $\mathcal{H}_X$ and $\mathcal{H}_Z$ infinite-dimensional and separable. We will show that the testing problem against Q is hard for any $Q \in \mathcal{Q}$. In particular, this includes the typical functional data setting where $\mathcal{H}_Z = L^2([0,1], \mathbb{R})$. It follows that the Gaussian conditional independence problem is hard in the same settings when the null distributions are not restricted to match the marginals of Q.

Preliminary results

In this section, we consider finite-dimensional $\mathcal{H}_X$ and infinite-dimensional $\mathcal{H}_Z$. We will need a lemma using the theory of conditional Hilbertian Gaussian distributions from Section 2.7.3.

Lemma 2.4. *Let (X, Y, Z) be jointly Gaussian on $\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y} \times \mathcal{H}$ and assume that the covariance operator of Z is injective. Then there exists a basis $(e_k)_{k\in\mathbb{N}}$ of $\mathcal{H}$ such that*

$$(X, Y) \perp\!\!\!\perp Z_{d_X+d_Y+1}, \ldots \mid Z_1, \ldots Z_{d_X}, Z_{d_X+1}, \ldots, Z_{d_X+d_Y}$$

where $Z_k := \langle Z, e_k\rangle$.

Proof. Note that $\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y} \times \mathcal{H}_Z$ is itself a Hilbert space and decompose it as $(\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}) \oplus \mathcal{H}_Z$. Let $\mathscr{C}_Z := \mathrm{Cov}(Z)$, $\mathscr{C}_{(X,Y)} := \mathrm{Cov}((X, Y))$ (the covariance of the joint vector (X, Y)) and $\mathscr{C}_{(X,Y),Z} := \mathrm{Cov}((X, Y), Z)$. We can apply Proposition 2.7 to see that (X, Y) conditional on Z is Gaussian with mean $\mathscr{C}_{(X,Y),Z}\mathscr{C}_Z^\dagger Z$ and covariance operator $\mathscr{C}_{(X,Y)} - \mathscr{C}_{(X,Y),Z}\mathscr{C}_Z^\dagger\mathscr{C}_{(X,Y),Z}^*$. The operator $\mathscr{A} := \mathscr{C}_{(X,Y),Z}\mathscr{C}_Z^\dagger$ maps from $\mathcal{H}$ to $\mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}$ and thus is at most a rank $d_X + d_Y$ operator. By Hsing and Eubank (2015, Theorem 3.3.7 7.) this implies that the rank of $\mathscr{A}^*$ is also at most $d_X + d_Y$. Furthermore, Hsing and Eubank (2015, Theorem 3.3.7 6.) yields that $\mathcal{H} = \mathrm{Ker}(\mathscr{A}) \oplus \mathrm{Im}(\mathscr{A}^*)$. Using this decomposition we can write $Z = (Z_{\mathrm{Ker}(\mathscr{A})}, Z_{\mathrm{Im}(\mathscr{A}^*)})$ and note that by construction $\mathscr{A}Z = \mathscr{A}Z_{\mathrm{Im}(\mathscr{A}^*)}$ thus the conditional distribution of (X, Y) given Z only depends on $Z_{\mathrm{Im}(\mathscr{A}^*)}$. In total, we have shown by Proposition 2.4 that $(X, Y) \perp\!\!\!\perp Z_{\mathrm{Ker}(\mathscr{A})} \mid Z_{\mathrm{Im}(\mathscr{A}^*)}$. Letting r denote the rank of $\mathscr{A}^*$, if we start with a basis for $\mathrm{Im}(\mathscr{A}^*)$ and append vectors to form a basis for $\mathcal{H}$ using the Gram–Schmidt procedure, we get a basis where

$$(X, Y) \perp\!\!\!\perp Z_{r+1}, \ldots \mid Z_1, \ldots Z_r.$$

Since $r \leq d_X + d_Y$, the weak union property of conditional independence yields

$$(X, Y) \perp\!\!\!\perp Z_{d_X+d_Y+1}, \ldots \mid Z_1, \ldots Z_{d_X}, Z_{d_X+1}, \ldots, Z_{d_Y+d_X},$$

as desired. □

Using this lemma and Lemma 2.3 and Theorem 2.7 from the previous section, we can prove the hardness result for finite-dimensional $\mathcal{H}_X$ and $\mathcal{H}_Y$.

Theorem 2.8. *Let $Q \in \mathcal{Q}$ be a distribution consisting of n i.i.d. copies of jointly Gaussian and injective (X, Y, Z) on $(\mathbb{R}^{d_X}, \mathbb{R}^{d_Y}, \mathcal{H}_Z)$ for $d_X, d_Y \in \mathbb{N}$ and any infinite-dimensional and separable $\mathcal{H}_Z$. Consider the testing problem described in Section 2.2 with $\mathcal{H}_X = \mathbb{R}^{d_X}$ and $\mathcal{H}_Z$ as above and let ψ be the test function of a size α test over $\mathcal{P}_0^Q$. Then ψ has power at most α against Q.*

Proof. Assume for contradiction that ψ is a test of size α over $\mathcal{P}_0^{\mathcal{Q}}$ with power $\alpha + \varepsilon$ for some $\varepsilon > 0$ against Q. Let (X, Y, Z) be distributed as one of the n i.i.d. copies constituting Q. By Lemma 2.4, we can express Z in a basis $(e_k)_{k\in\mathbb{N}}$ such that defining $Z_k = \langle Z, e_k\rangle$, we have

$$(X, Y) \perp\!\!\!\perp Z_{d_X+d_Y+1}, \ldots \mid Z_1, \ldots, Z_{d_X}, Z_{d_X+1}, \ldots, Z_{d_X+d_Y}.$$

By the weak union property of conditional independence, this implies that

$$(X, Y) \perp\!\!\!\perp Z_{d+1}, \dots \mid Z_1, \dots, Z_d$$

for any $d \geq d_X + d_Y$.

Choose now an arbitrary $d \geq d_X + d_Y$ and let $\tilde{Q}$ denote the distribution of n i.i.d. copies of $(X, Y, Z_1, \dots, Z_d)$ under Q. Consider the testing problem described in Section 2.2 with $\mathcal{H}_X = \mathbb{R}^{d_X}$ and $\mathcal{H}_Z = \mathbb{R}^d$. We can construct a test in this setting by defining new observations $(\check{X}, \check{Y}, \check{Z})$ with values in $(\mathbb{R}^{d_X}, \mathbb{R}^{d_Y}, \mathcal{H}_Z)$ and applying ψ. We form the new observations by setting $\check{X} := \tilde{X}$, $\check{Y} := \tilde{Y}$ and $\check{Z} := (\tilde{Z}_1, \dots, \tilde{Z}_d, Z^\circ_{d+1}, Z^\circ_{d+2}, \dots)$, where $Z^\circ_{d+1}, Z^\circ_{d+2}, \dots$ are sampled from the conditional distribution $Z_{d+1}, Z_{d+2}, \dots \mid Z_1 = \tilde{Z}_1, \dots Z_d = \tilde{Z}_d$. If the original sample is from a distribution in $\mathcal{P}_0^{\tilde{Q}}$ then the modified sample will be from a null distribution in $\mathcal{P}_0^{Q}$, thus the test has size α over $\mathcal{P}_0^{\tilde{Q}}$. Similarly, if $(\tilde{X}, \tilde{Y}, \tilde{Z}) \sim \tilde{Q}$, the modified sample will have distribution Q and hence the test has power $\alpha + \varepsilon$ against $\tilde{Q}$.

By Lemma 2.3 this implies the existence of a $d_X \times d_Y$ block-diagonal matrix R with diagonal entries in the open unit interval, a Gaussian distribution Q' on $(\mathbb{R}^{d_X}, \mathbb{R}^{d_Y}, \mathbb{R}^d)$ where if $(X', Y', Z') \sim Q'$, X', Y' and Z' are mean zero with identity covariance matrix, $\mathrm{Cov}(X', Z') = \mathrm{Cov}(Y', Z') = 0$ and $\mathrm{Cov}(X', Y') = R$, and a test with size α over $\mathcal{P}_0^{Q'}$ with power $\alpha + \varepsilon$ against Q'. Since d was arbitrary, this contradicts Theorem 2.7. □

Proofs of Theorem 2.1 and Proposition 2.1

In this section we prove Theorem 2.1 and Proposition 2.1. We do this by extending the results from the previous section to the situation where at most one of X and Y are infinite-dimensional.

Lemma 2.5. *Let (X, Y, Z) be jointly Gaussian on $\mathbb{R}^{d_X} \times \mathcal{H}_Y \times \mathcal{H}_Z$ and assume that the covariance operator of (Y, Z) is injective. Then there exists a basis $(e_k)_{k\in\mathbb{N}}$ of $\mathcal{H}_Y$ such that*

$$X \perp\!\!\!\perp Y_{d_X+1}, \dots \mid Y_1, \dots, Y_{d_X}, Z$$

where $Y_k := \langle Y, e_k \rangle$.

Proof. Note that $\mathbb{R}^{d_X} \times \mathcal{H}_Y \times \mathcal{H}_Z$ is again a Hilbert space and decompose it as $\mathbb{R}^{d_X} \oplus (\mathcal{H}_Y \times \mathcal{H}_Z)$. Let $\mathscr{C}_{(Y,Z)} := \mathrm{Cov}((Y, Z))$ (the covariance of the joint vector (Y, Z)), $\mathscr{C}_X := \mathrm{Cov}(X)$ and $\mathscr{C}_{X,(Y,Z)} := \mathrm{Cov}(X, (Y, Z))$. We can apply Proposition 2.7 to see that X conditional on (Y, Z) is Gaussian with mean $\mathscr{C}_{X,(Y,Z)}\mathscr{C}^\dagger_{(Y,Z)}(Y, Z)$ and covariance operator $\mathscr{C}_X - \mathscr{C}_{X,(Y,Z)}\mathscr{C}^\dagger_{(Y,Z)}\mathscr{C}^*_{X,(Y,Z)}$. The operator $\mathscr{A} = \mathscr{C}_{X,(Y,Z)}\mathscr{C}^\dagger_{(Y,Z)}$ maps from $\mathcal{H}_Y \times \mathcal{H}_Z$ to $\mathbb{R}^{d_X}$ and thus is at most a rank d_X operator. By Hsing and Eubank (2015, Theorem 3.3.7 7.) this implies that the rank of $\mathscr{A}^*$ is also at most d_X. Furthermore, Hsing and Eubank (2015, Theorem 3.3.7 6.) yields that $\mathcal{H}_Y \times \mathcal{H}_Z = \mathrm{Ker}(\mathscr{A}) \oplus \mathrm{Im}(\mathscr{A}^*)$.

Using this decomposition we can write $(Y, Z) = ((Y, Z)_{\mathrm{Ker}(\mathscr{A})}, (Y, Z)_{\mathrm{Im}(\mathscr{A}^*)})$ and note that by construction $\mathscr{A}(Y, Z) = \mathscr{A}(Y, Z)_{\mathrm{Im}(\mathscr{A}^*)}$ thus the conditional distribution of X given (Y, Z) only depends on $(Y, Z)_{\mathrm{Im}(\mathscr{A}^*)}$. In total, we have shown by Proposition 2.4 that

$X \perp\!\!\!\perp (Y,Z)_{\mathrm{Ker}(\mathscr{A})} \,|\, (Y,Z)_{\mathrm{Im}(\mathscr{A}^*)}$ which implies by the weak union property of conditional independence that $X \perp\!\!\!\perp Y_{\mathrm{Ker}(\mathscr{A})} \,|\, Y_{\mathrm{Im}(\mathscr{A}^*)}, Z$.

Any basis of $\mathrm{Im}(\mathscr{A}^*)$ will consist of at most d_X elements. Forming the span of the $\mathcal{H}_Y$-components of the basis vectors will yield a subspace of $\mathcal{H}_Y$ that contains the projection onto $\mathcal{H}_Y$ of $\mathrm{Im}(\mathscr{A}^*)$. Thus, letting r denote the rank of $\mathscr{A}^*$, we can append vectors and form a basis for $\mathcal{H}_Y$ using the Gram–Schmidt procedure to get a basis where

$$X \perp\!\!\!\perp Y_{r+1}, \dots \,|\, Y_1, \dots, Y_r, Z.$$

Since $r \leq d_X$, the weak union property of conditional independence yields

$$X \perp\!\!\!\perp Y_{d_X+1}, \dots \,|\, Y_1, \dots, Y_{d_X}, Z$$

as desired. □

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1. Assume without loss of generality that $\mathcal{H}_X$ is finite-dimensional and thus $\mathcal{H}_X$ is isomorphic to a real vector space, and we will instead denote $\mathcal{H}_X = \mathbb{R}^{d_X}$ where d_X is the dimension of $\mathcal{H}_X$.

Assume for contradiction that ψ is a test of size α, with power $\alpha + \epsilon$ for some $\epsilon > 0$ against Q. Let (X, Y, Z) be distributed as one of the n i.i.d. copies constituting Q. By Lemma 2.5 we can express Y in a basis $(e_k)_{k\in\mathbb{N}}$ such that defining $Y_k = \langle Y, e_k \rangle$, we have

$$X \perp\!\!\!\perp Y_{d_X+1}, \dots \,|\, Y_1, \dots, Y_{d_X}, Z.$$

By the weak union property of conditional independence, this implies that

$$X \perp\!\!\!\perp Y_{d+1}, \dots \,|\, Y_1, \dots, Y_d, Z$$

for any $d \geq d_X$.

Choose now an arbitrary $d \geq d_X + d_Y$ and let $\tilde{Q}$ denote the distribution of n i.i.d. copies of $(X, Y_1, \dots, Y_d, Z)$ under Q. Consider the testing problem described in Section 2.2 with $\mathcal{H}_X = \mathbb{R}^{d_X}$ and $\mathcal{H}_Z$ as above. We can construct a test in this setting by defining new observations $(\check{X}, \check{Y}, \check{Z})$ with values in $(\mathbb{R}^{d_X}, \mathcal{H}_Y, \mathcal{H}_Z)$ and applying ψ. We form the new observations by setting $\check{X} := \tilde{X}$, $\check{Z} := \tilde{Z}$ and $\check{Y} := (\tilde{Y}_1, \dots, \tilde{Y}_d, Y^\circ_{d+1}, Y^\circ_{d+2}, \dots)$, where $Y^\circ_{d+1}, Y^\circ_{d+2}, \dots$ are sampled from the conditional distribution $Y_{d+1}, Y_{d+2}, \dots \,|\, Y_1 = \tilde{Y}_1, \dots, Y_d = \tilde{Y}_d, Z = \tilde{Z}$. If the original sample is from a distribution in $\mathcal{P}_0^{\tilde{Q}}$, then the modified sample will be from a null distribution in $\mathcal{P}_0^{Q}$, thus the test has size α over $\mathcal{P}_0^{\tilde{Q}}$. Similarly, if $(\tilde{X}, \tilde{Y}, \tilde{Z}) \sim \tilde{Q}$, the modified sample will have distribution Q and hence the test has power $\alpha + \epsilon$ against the distribution of $(X, Y_1, \dots, Y_d, Z)$. But this contradicts Theorem 2.8. □

A similar strategy can be employed to prove Proposition 2.1.

Proof of Proposition 2.1. We can repeat the arguments of Theorem 2.8 and Theorem 2.1 without using Lemma 2.4 and Lemma 2.5 since we can use the basis $(e_k)_{k\in\mathbb{N}}$ instead. □

2.7.3 Auxiliary results about conditional distributions on Hilbert spaces

Let us first recall how to formally define a conditional distribution. We follow Dudley (2002, Chapter 10.2) and Rønn-Nielsen and Hansen (2014).

Definition 2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{D}$ be a sub-σ-algebra of $\mathcal{F}$ and let $\mathbb{P}_{|\mathcal{D}}$ denote the restriction of $\mathbb{P}$ to $\mathcal{D}$. Let X be a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$ mapping into a measurable space $(\mathcal{X}, \mathcal{A})$. We say that a function $P_{X\,|\,\mathcal{D}} : \mathcal{A} \times \Omega \to [0,1]$ is a* conditional distribution for X given $\mathcal{D}$ *if the following two conditions hold.*

(i) For each $A \in \mathcal{A}$, $P_{X\,|\,\mathcal{D}}(A, \cdot) = \mathbb{E}(\mathbb{1}_{\{(X\in A)\}}\,|\,\mathcal{D}) = \mathbb{P}(X \in A\,|\,\mathcal{D})$ $\mathbb{P}_{|\mathcal{D}}$-a.s.

(ii) For $P_{|\mathcal{D}}$ almost every $\omega \in \Omega$, $P_{X\,|\,\mathcal{D}}(\cdot, \omega)$ is a probability measure on $(\mathcal{X}, \mathcal{A})$.

We are mainly interested in conditioning on the value of some random variable which leads to the following definition.

Definition 2.2. *Consider random variables X and Y defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in the measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{G})$, respectively. We say that a function $P_{Y\,|\,X} : \mathcal{G} \times \mathcal{X} \to [0,1]$ is a* conditional distribution for Y given X *if the following conditions hold.*

(i) For each $x \in \mathcal{X}$, $P_{Y\,|\,X}(\cdot, x)$ is a probability measure on $(\mathcal{Y}, \mathcal{G})$.

(ii) For each $G \in \mathcal{G}$, $P_{Y\,|\,X}(G, \cdot)$ is $\mathcal{A}$-$\mathbb{B}$ measurable, where $\mathbb{B}$ denotes the Borel σ-algebra on $\mathbb{R}$.

(iii) For each $A \in \mathcal{A}$

$$\mathbb{P}(X \in A, Y \in G) = \int_{(X\in A)} P_{Y\,|\,X}(G, X(\omega))\,\mathrm{d}\mathbb{P}(\omega) = \int_A P_{Y\,|\,X}(G, x)\,\mathrm{d}X(\mathbb{P})(x),$$

where $X(\mathbb{P})$ is the push-forward measure of X under $\mathbb{P}$, i.e. the measure on $(\mathcal{X}, \mathcal{A})$ such that $X(\mathbb{P})(A) = \mathbb{P}(X \in A)$ for $A \in \mathcal{A}$.

Informally, we write $Y\,|\,X$ for the conditional distribution of Y given X and $Y\,|\,X = x$ for the measure $P_{Y\,|\,X}(\cdot, x)$. If a function $Q : \mathcal{G} \times \mathcal{X} \to [0,1]$ only satisfies the first two conditions, we say that Q is a $(\mathcal{X}, \mathcal{A})$-Markov kernel on $(\mathcal{Y}, \mathcal{G})$.

The connection between the previous two definitions can be seen by viewing X and Y as random variables on the probability space $(\mathcal{X}\times\mathcal{Y}, \mathcal{A}\otimes\mathcal{G}, (X,Y)(\mathbb{P}))$ where $(X,Y)(\mathbb{P})$ is the joint push-forward measure of X and Y under $\mathbb{P}$. If we then let $\mathcal{D}$ be the smallest σ-algebra making the projection onto the $\mathcal{X}$-space measurable, we see by letting $P_{Y\,|\,\mathcal{D}}(G, (x,y)) = P_{Y\,|\,X}(G, x)$ that $P_{Y\,|\,X}$ also satisfies the conditions of the first definition. For more on this perspective, see Dudley (2002, Theorem 10.2.1). It is non-trivial to show the existence of conditional distributions, however, we do have the following result from Dudley (2002, Theorem 10.2.2).

Lemma 2.6. *Consider random variables X and Y defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in the measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{G})$ respectively. If $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces and $\mathcal{A}$ and $\mathcal{G}$ are their respective Borel σ-algebras then the conditional distribution for Y given X exists.*

We will consider real-valued and Hilbertian random variables in the following, thus we are free to assume the existence of conditional distributions wherever needed. Before we delve into the main preliminary results about Hilbertian conditional distributions, we present some fundamental results from the theory of regular conditional distributions. For measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{G})$, we let $i_x : \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ denote the inclusion map, i.e. $i_x(y) = (x, y)$. This is a $\mathcal{G} - \mathcal{A} \otimes \mathcal{G}$ measurable mapping for each fixed x. The following four results are included for completeness and can be found in Rønn-Nielsen and Hansen (2014, Lemma 1.1.4, Theorem 1.2.1, Theorem 2.1.1 & Theorem 3.5.5). Unless otherwise specified, for these results X, Y and Z are random variables on measurable spaces $(\mathcal{X}, \mathcal{A})$, $(\mathcal{Y}, \mathcal{G})$ and $(\mathcal{Z}, \mathcal{K})$ respectively.

Lemma 2.7. *Let Q be a $(\mathcal{X}, \mathcal{A})$-Markov kernel on $(\mathcal{Y}, \mathcal{G})$ and let $\mathbb{B}$ denote the Borel σ-algebra on $\mathbb{R}$. For each $C \in \mathcal{A} \otimes \mathcal{G}$ the map*

$$x \mapsto Q(i_x^{-1}(C), x)$$

is $\mathcal{A}$-$\mathbb{B}$ measurable.

Proof. Let

$$\mathcal{D} = \{C \in \mathcal{A} \otimes \mathcal{G} \,|\, x \mapsto Q(i_x^{-1}(C), x) \text{ is } \mathcal{A}\text{-}\mathbb{B} \text{ measurable}\}$$

and consider a product set $A \times G \in \mathcal{A} \otimes \mathcal{G}$. Clearly,

$$i_x^{-1}(A \times G) = \begin{cases} \emptyset & \text{if } x \notin A \\ B & \text{if } x \in A \end{cases}$$

and therefore

$$Q(i_x^{-1}(A \times G), x) = \begin{cases} 0 & \text{if } x \notin A \\ Q(G, x) & \text{if } x \in A \end{cases} = \mathbb{1}_A(x) Q(G, x).$$

This is a product of two $\mathcal{A}$-$\mathbb{B}$ measurable functions and is thus also $\mathcal{A}$-$\mathbb{B}$ measurable. This shows that $\mathcal{D}$ contains all product sets and since the product sets are an intersection-stable generator of $\mathcal{A} \otimes \mathcal{G}$, we are done if we can show that $\mathcal{D}$ is a Dynkin class by Schilling (2017, Theorem 5.5).

We have already shown that product sets are in $\mathcal{D}$ which includes $\mathcal{X} \times \mathcal{Y}$. If $C_1, C_2 \in \mathcal{D}$ where $C_1 \subseteq C_2$ then clearly also $i_x^{-1}(C_1) \subseteq i_x^{-1}(C_2)$ and further $i_x^{-1}(C_2 \setminus C_1) = i_x^{-1}(C_2) \setminus i_x^{-1}(C_1)$. This implies that

$$Q(i_x^{-1}(C_2 \setminus C_1), x) = Q(i_x^{-1}(C_2), x) - Q(i_x^{-1}(C_1), x)$$

which is the difference of two $\mathcal{A}$-$\mathbb{B}$ measurable functions and is thus also $\mathcal{A}$-$\mathbb{B}$ measurable. Hence, $C_2 \setminus C_1 \in \mathcal{D}$. Finally, assume that $C_1 \subseteq C_2 \subseteq \cdots$ is an increasing sequence of $\mathcal{D}$-sets. Similarly

to above we have $i_x^{-1}(C_1) \subseteq i_x^{-1}(C_2) \subseteq \cdots$ and

$$i_x^{-1}\left(\bigcup_{n=1}^{\infty} C_n\right) = \bigcup_{n=1}^{\infty} i_x^{-1}\left(C_n\right).$$

Then

$$Q\left(i_x^{-1}\left(\bigcup_{n=1}^{\infty} C_n\right), x\right) = Q\left(\bigcup_{n=1}^{\infty} i_x^{-1}\left(C_n\right), x\right) = \lim_{n\to\infty} Q(i_x^{-1}(C_n), x).$$

The limit is $\mathcal{A}$-$\mathbb{B}$ measurable since each of the functions $x \mapsto Q(i_x^{-1}(C_n), x)$ are measurable. Hence, $\mathcal{D}$ is a Dynkin class, and we have the desired result. □

Proposition 2.2. *Let μ be a probability measure on $(\mathcal{X}, \mathcal{A})$ and let Q be a $(\mathcal{X}, \mathcal{A})$-Markov kernel on $(\mathcal{Y}, \mathcal{G})$. There exists a uniquely determined probability measure λ on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{G})$ satisfying*

$$\lambda(A \times G) = \int_A Q(G, x), \mathrm{d}\mu(x)$$

for all $A \in \mathcal{A}$ and $G \in \mathcal{G}$. Furthermore, for $C \in \mathcal{A} \otimes \mathcal{G}$

$$\lambda(C) = \int Q(i_x^{-1}(C), x), \mathrm{d}\mu(x).$$

Proof. Uniqueness follows from Schilling (2017, Theorem 5.7) since λ is determined on the product sets which form an intersection-stable generator of $\mathcal{A} \otimes \mathcal{G}$.

For existence, we show that λ as defined for general $C \in \mathcal{A} \otimes \mathcal{G}$ is a measure. The integrand is measurable by Lemma 2.7 and since Q is non-negative, the integral is well-defined with values in $[0, \infty]$. Let $C_1, C_2 \ldots$ be a sequence of disjoint sets in $\mathcal{A} \otimes \mathcal{G}$. Then for each $x \in \mathcal{X}$ the sets $i_x^{-1}(C_1), i_x^{-1}(C_2), \ldots$ are disjoint as well. Hence,

$$\begin{aligned}\lambda\left(\bigcup_{n=1}^{\infty} C_n\right) &= \int Q\left(i_x^{-1}\left(\bigcup_{n=1}^{\infty} C_n\right), x\right) \mathrm{d}\mu(x) = \int \sum_{n=1}^{\infty} Q\left(i_x^{-1}(C_n), x\right) \mathrm{d}\mu(x) \\ &= \sum_{n=1}^{\infty} \int Q\left(i_x^{-1}(C_n), x\right) \mathrm{d}\mu(x) = \sum_{n=1}^{\infty} \lambda(C_n)\end{aligned}$$

where the second equality uses that $Q(\cdot, x)$ is a measure and the third uses monotone convergence to interchange integration and summation. Since also

$$\lambda(\mathcal{X} \times \mathcal{Y}) = \int Q(i_x^{-1}(\mathcal{X} \times \mathcal{Y}), x) \, \mathrm{d}\mu(x) = \int Q(\mathcal{Y}, x) \, \mathrm{d}\mu(x) = \int 1 \, \mathrm{d}\mu(x) = 1$$

λ is a probability measure, and it follows that

$$\lambda(A \times G) = \int Q(i_x^{-1}(A \times G), x) \, \mathrm{d}\mu(x) = \int_A Q(G, x) \, \mathrm{d}\mu(x)$$

for all $A \in \mathcal{A}$ and $G \in \mathcal{G}$ as desired. □

Proposition 2.3. *Assume that $P_{Y\,|\,X}$ is the conditional distribution of Y given X. Let $(\mathcal{Z}, \mathcal{K})$ be another measurable space and let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a measurable mapping. Define $Z = \phi(X, Y)$. Then the conditional distribution of Z given X exists and for $K \in \mathcal{K}$ and $x \in \mathcal{X}$ is given by*

$$P_{Z\,|\,X}(K, x) = P_{Y\,|\,X}((\phi \circ i_x)^{-1}(K), x).$$

Proof. Clearly $P_{Z\,|\,X}(\cdot, x)$ is a probability measure for every $x \in \mathcal{X}$ and Lemma 2.7 yields that $P_{Z\,|\,X}(K, \cdot)$ is $\mathcal{A}$-$\mathbb{B}$ measurable for every $K \in \mathcal{K}$. It remains to show that $P_{Z\,|\,X}$ satisfies the third condition required to be the conditional distribution of Z given X. For $A \in \mathcal{A}$ and $K \in \mathcal{K}$ we get that

$$\mathbb{P}(X \in A, Z \in K) = \mathbb{P}((X, Y) \in (A \times \mathcal{Y}) \cap \phi^{-1}(K))$$

and hence by Proposition 2.2, we get that

$$\mathbb{P}(X \in A, Z \in K) = \int P_{Y\,|\,X}(i_x^{-1}((A \times \mathcal{Y}) \cap \phi^{-1}(K)), x)\, \mathrm{d}X(\mathbb{P})(x).$$

Since

$$i_x^{-1}((A \times \mathcal{Y}) \cap \phi^{-1}(K)) = \begin{cases} \emptyset & \text{if } x \notin A \\ i_x^{-1}(\phi^{-1}(K)) & \text{if } x \in A \end{cases},$$

we get

$$\mathbb{P}(X \in A, Z \in K) = \int_A P_{Y\,|\,X}(i_x^{-1}(\phi^{-1}(K)), x)\, \mathrm{d}X(\mathbb{P})(x) = \int_A P_{Z\,|\,X}(K, x)\, \mathrm{d}X(\mathbb{P})(x),$$

proving the desired result. □

Proposition 2.4. *Suppose that conditional distribution $P_{Y\,|\,(X,Z)}$ of Y given (X, Z) has the structure*

$$P_{Y\,|\,(X,Z)}(G, (x, z)) = Q(G, z)$$

for some $Q : \mathcal{G} \times \mathcal{Z}$ where for every $z \in \mathcal{Z}$, $Q(\cdot, z)$ is a probability measure. Then Q is a Markov kernel, Q is the conditional distribution of Y given Z and $X \perp\!\!\!\perp Y \,|\, Z$.

Proof. That Q is a Markov kernel follows immediately from the fact that $P_{Y\,|\,(X,Z)}$ is a Markov kernel. To see that Q is the conditional distribution of Y given Z, note that defining $\pi_Z : \mathcal{X} \times \mathcal{Z} \to \mathcal{Z}$ to be the projection onto $\mathcal{Z}$, we get

$$\begin{aligned} \mathbb{P}(Z \in K, Y \in G) &= \mathbb{P}((X, Z) \in \pi_Z^{-1}(K), Y \in G) \\ &= \int_{\pi_Z^{-1}(K)} P_{Y\,|\,(X,Z)}(G, (x, z))\, \mathrm{d}(X, Z)(P)(x, z) \\ &= \int_{\pi_Z^{-1}(K)} Q(G, \pi_Z(x, z))\, \mathrm{d}(X, Z)(P)(x, z) = \int_K Q(G, z)\, \mathrm{d}Z(P)(z), \end{aligned}$$

by viewing $Z(P)$ as the image measure of $(X, Z)(P)$ under π_Z and applying Schilling (2017, Theorem 14.1). For every $G \in \mathcal{G}$, $Q(G, Z)$ is a version of the conditional probability $\mathbb{P}(Y \in$

$G \mid Z) = \mathbb{E}(1_{(Y \in G)} \mid Z)$ since $Q(G, Z)$ is clearly measurable with respect to $\sigma(Z)$ and

$$\int_{(Z \in K)} 1_{(Y \in G)} \, \mathrm{d}P = \mathbb{P}(Z \in K, Y \in G) = \int_{(Z \in K)} Q(G, Z) \, \mathrm{d}P.$$

The same argument applies to show that $P_{Y \mid (X,Z)}(G, (X, Z))$ is a version of $\mathbb{P}(Y \in G \mid X, Z)$. Hence, for every $G \in \mathcal{G}$

$$\mathbb{P}(Y \in G \mid Z) = Q(G, Z) = P_{Y \mid (X,Z)}(G, (X, Z)) = \mathbb{P}(Y \mid X, Z)$$

and thus $X \perp\!\!\!\perp Y \mid Z$ as desired. □

With these results we are ready to start considering Hilbertian conditional distributions.

Remark 2.1. *In the following we will repeatedly consider orthogonal decompositions of Hilbert spaces. We write $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ if every $h \in \mathcal{H}$ can be written as $h = h_1 + h_2$ where $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ and $\mathcal{H}_1 \perp \mathcal{H}_2$. If an operator $\mathscr{A}$ is defined on $\mathcal{H}$, the decomposition induces four operators: $\mathscr{A}_{11}$ and $\mathscr{A}_{21}$, the $\mathcal{H}_1$ and $\mathcal{H}_2$ components of the restriction of $\mathscr{A}$ to $\mathcal{H}_1$ and similarly $\mathscr{A}_{12}$ and $\mathscr{A}_{22}$, the $\mathcal{H}_1$ and $\mathcal{H}_2$ components of the restriction of $\mathscr{A}$ to $\mathcal{H}_2$. We can write $\mathscr{A}$ as the sum of these four operators. If X is a random variable on $\mathcal{H}$ and $\mathcal{H}_1$ and $\mathcal{H}_2$ are as above, we can similarly decompose X into (X_1, X_2) where $X_1 \in \mathcal{H}_1$ and $X_2 \in \mathcal{H}_2$. If $\mathscr{C}$ is the covariance operator of X, we can decompose it as mentioned above and, in particular, we have $\mathscr{C}_{11} = \mathrm{Cov}(X_1)$, $\mathscr{C}_{22} = \mathrm{Cov}(X_2)$ and $\mathscr{C}_{12} = \mathscr{C}_{21}^* = \mathrm{Cov}(X_1, X_2)$, where $\mathscr{C}_{21}^*$ denotes the adjoint of $\mathscr{C}_{21}$. This is analogous to the usual block matrix decomposition of the covariance matrix of multivariate random variables.*

We will need two results that are fundamental in the theory of the multivariate Gaussian distribution.

Proposition 2.5. *Let X be Gaussian on $\mathcal{H}$ and assume that $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$. Define (X_1, X_2) to be the corresponding decomposition of X. Then $X_1 \perp\!\!\!\perp X_2$ if and only if $\mathrm{Cov}(X_1, X_2) = 0$.*

Proof. We show that $\mathrm{Cov}(X_1, X_2) = 0$ implies independence since the other direction is trivial. We will use the approach of characteristic functionals as described in detail in Vakhania et al. (1987, Chapter IV). The characteristic functional of a random variable (technically, the distribution of the random variable) is the mapping defined on $\mathcal{H}$ where $h \mapsto \mathbb{E}[\exp(i\langle X, h\rangle)]$. Vakhania et al. (1987, Theorem IV.2.4) state that for Gaussian X with mean μ and covariance operator $\mathscr{C}$ the characteristic functional is

$$\phi_X(h) = \exp\left(i\langle \mu, h\rangle - \frac{1}{2}\langle \mathscr{C}h, h\rangle\right).$$

Vakhania et al. (1987, Chapter IV, Proposition 2.2 + Corollary) state that X_1 and X_2 are independent if the characteristic functional of X factorises into the product of their respective characteristic functionals. By the assumption that $\mathscr{C}_{12} = \mathrm{Cov}(X_1, X_2) = 0$, we can write the covariance as $\mathscr{C} = \mathscr{C}_1 + \mathscr{C}_2$ where $\mathscr{C}_i$ is the covariance of X_i. The result then follows by factorising the characteristic functional appropriately. □

Proposition 2.6. *Let X be Gaussian on $\mathcal{H}_1$ with mean μ and covariance operator $\mathscr{C}$ and let $\mathscr{A}$ be a bounded linear operator from $\mathcal{H}_1$ to $\mathcal{H}_2$ and $z \in \mathcal{H}_2$. Then $Y = \mathscr{A}X + z$ is Gaussian on $\mathcal{H}_2$ with mean $\mathscr{A}\mu + z$ and covariance operator $\mathscr{A}\mathscr{C}\mathscr{A}^*$ where $\mathscr{A}^*$ is the adjoint of $\mathscr{A}$.*

Proof. Throughout, we let $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ denote the inner products of $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively. By definition, for every $h_1 \in \mathcal{H}_1$, $\langle X, h_1 \rangle$ is Gaussian on $\mathbb{R}$. For every $h_2 \in \mathcal{H}_2$ we have

$$\langle Y, h_2 \rangle_2 = \langle \mathscr{A}X, h_2 \rangle_2 + \langle z, h_2 \rangle_2 = \langle X, \mathscr{A}^* h_2 \rangle_1 + \langle z, h_2 \rangle_2$$

thus Y is also Gaussian. Using the interchangeability of the Bochner integral and linear operators (see Hsing and Eubank (2015, Theorem 3.1.7)), we get the mean of Y immediately. By noting that for any $h, k \in \mathcal{H}_1$, we have

$$(\mathscr{A}h) \otimes k = \langle \mathscr{A}h, \cdot \rangle_2 k = \langle h, \mathscr{A}^* \cdot \rangle_1 k = (h \otimes k)\mathscr{A}^*,$$

the covariance result then follows by the same argument as for the mean. □

With these results we can now show that conditioning on an injective part of a Gaussian distribution on a Hilbert space yields another Gaussian distribution with mean and covariance given by the Hilbertian analogue of the well-known Gaussian conditioning formula.

Proposition 2.7. *Let X be mean zero Gaussian on $\mathcal{H}$ with covariance operator $\mathscr{C}$ and assume that $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$. Let (X_1, X_2) denote the corresponding decomposition of X. As discussed in Remark 2.1, we then set $\mathscr{C}_{11} := \mathrm{Cov}(X_1)$, $\mathscr{C}_{22} := \mathrm{Cov}(X_2)$ and $\mathscr{C}_{12} = \mathscr{C}_{21}^* := \mathrm{Cov}(X_1, X_2)$, where $\mathscr{C}_{21}^*$ denotes the adjoint of $\mathscr{C}_{21}$. If $\mathscr{C}_{22}$ is injective, i.e.*

$$Ker(\mathscr{C}_{22}) = \{h \in \mathcal{H}_2 \,|\, \mathscr{C}_{22}h = 0\} = \{0\}$$

then the conditional distribution of X_1 given X_2 is Gaussian on $\mathcal{H}_1$ with

$$\mathbb{E}(X_1 \,|\, X_2) = \mathscr{C}_{12}\mathscr{C}_{22}^{\dagger} X_2$$

and

$$\mathrm{Cov}(X_1 \,|\, X_2) = \mathscr{C}_{11} - \mathscr{C}_{12}\mathscr{C}_{22}^{\dagger}\mathscr{C}_{21},$$

where $\mathscr{C}_{22}^{\dagger}$ is the generalised inverse (or Moore–Penrose inverse) of $\mathscr{C}_{22}$.

Proof. Define $Z := X_1 - \mathscr{C}_{12}\mathscr{C}_{22}^{\dagger} X_2$. Note that since (Z, X_2) is a bounded linear transformation of (X_1, X_2), (Z, X_2) must be jointly Gaussian by Proposition 2.6. By Proposition 2.5, Z and X_2 are independent if $\mathrm{Cov}(Z, X_2) = 0$. We calculate the covariance and get

$$\mathrm{Cov}(Z, X_2) = \mathscr{C}_{12} - \mathscr{C}_{12}\mathscr{C}_{22}^{\dagger}\mathscr{C}_{22} = 0$$

by Hsing and Eubank (2015, Theorem 3.5.8 (3.18)) since $\mathrm{Ker}(\mathscr{C}_{22}) = 0$. This implies that the conditional distribution of Z given X_2 is simply the distribution of Z. We can find the complete distribution of Z by calculating the mean and covariance of Z, since Z is Gaussian. We get by

Proposition 2.6,

$$\mathbb{E}(Z) = \mathbb{E}(X_1) - \mathscr{C}_{12}\mathscr{C}_{22}^{\dagger}\mathbb{E}(X_2) = 0$$

and

$$\mathrm{Cov}(Z) = \mathscr{C}_{11} - \mathscr{C}_{12}\mathscr{C}_{22}^{\dagger}\mathscr{C}_{21}.$$

By Proposition 2.3, since we can write $X_1 = Z + \mathscr{C}_{12}\mathscr{C}_{22}^{\dagger}X_2$, the conditional distribution of X_1 given X_2 is as desired. □

2.8 Uniform convergence of random variables

In this section we develop some background theory that will be useful when considering simultaneous convergence of sequences with varying distributions. In particular, we are interested the convergence of a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ defined on a measurable space $(\Omega, \mathcal{F})$ with a family of probability measures $(\mathbb{P}_\theta)_{\theta\in\Theta}$. For each $\theta \in \Theta$ the distribution of $(X_n)_{n\in\mathbb{N}}$ will change as the background measure $\mathbb{P}_\theta$ changes. We are also interested in the convergence of θ-dependent functions of X_n such as the conditional expectation with respect to $\mathbb{P}_\theta$ of X_n given a sub-σ-algebra $\mathcal{D}$ of $\mathcal{F}$. To allow for such considerations, the definitions given here will be more general than in Section 2.4 and will allow for a family of random variables $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ to converge to a family of random variables $(X_\theta)_{\theta\in\Theta}$.

The material in this section extends the work of Kasy (2019) and Bengs and Holzmann (2019) to Hilbertian and Banachian random variables and also adds further characterisations of their central assumptions for families of real-valued random variables.

Unless stated otherwise, we consider the following setup for the remainder of this section. Let $(\Omega, \mathcal{F})$ be a measurable space, $(\mathbb{P}_\theta)_{\theta\in\Theta}$ a family of probability measure on $(\Omega, \mathcal{F})$ where Θ is any set and $(\mathcal{B}, \mathbb{B}(\mathcal{B}))$ a separable Banach space with its Borel σ-algebra. Let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ and $(X_\theta)_{\theta\in\Theta}$ be families of random variables defined on $(\Omega, \mathcal{F})$ with values in $\mathcal{B}$. All additional random variables are also defined on $(\Omega, \mathcal{F})$. We write $\mathbb{E}_\theta$ for the expectation with respect to $\mathbb{P}_\theta$.

Definition 2.3 (Uniform convergence of random variables)**.** *(i) We say that* $X_{n,\theta}$ converges uniformly in distribution over Θ to X_θ and write $X_{n,\theta} \overset{\mathcal{D}}{\underset{\Theta}{\rightrightarrows}} X_\theta$ *if*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} d_{BL}^{\theta}(X_{n,\theta}, X_\theta) = 0,$$

where

$$d_{BL}^{\theta}(X_{n,\theta}, X_\theta) := \sup_{f\in BL_1} \left|\mathbb{E}_\theta(f(X_{n,\theta})) - \mathbb{E}_\theta(f(X_\theta))\right|,$$

and BL_1 *denotes the set of all functions* $f : \mathcal{B} \to [-1, 1]$ *that are Lipschitz with constant at most* 1*. When* Θ *is clear from the context, we simply write* $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$ *and say that* $X_{n,\theta}$ *converges uniformly in distribution to* X_θ*. When considering collections of random variables that do not depend on* θ *except through the measure on the domain of the random variables, we simply write* $X_n \overset{\mathcal{D}}{\rightrightarrows} X$*.*

(ii) We say that $X_{n,\theta}$ converges uniformly in probability over Θ to X_θ and write $X_{n,\theta} \overset{P}{\underset{\Theta}{\rightrightarrows}} X_\theta$ *if, for any* $\epsilon > 0$,

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \mathbb{P}_\theta(\|X_{n,\theta} - X_\theta\| \geq \epsilon) = 0.$$

We write $X_{n,\theta} \overset{P}{\rightrightarrows} X_\theta$ *and simply say that* $X_{n,\theta}$ *converges uniformly in probability to* X_θ *when* Θ *is clear from the context. When considering collections of random variables that do not depend on* θ *except through the measure on the domain of the random variables, we simply write* $X_n \overset{P}{\rightrightarrows} X$.

Using a slight abuse of notation, we write $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} 0$ and $X_{n,\theta} \overset{P}{\rightrightarrows} 0$ to mean that $X_{n,\theta}$ converges uniformly to the family of random variables X_θ that is equal to 0 for all $\omega \in \Omega$ and any $\theta \in \Theta$. Note that if $(\mathbb{P}_\theta)_{\theta\in\Theta}$ contains a single element, we recover the standard definitions of convergence in distribution and probability. We have the following helpful characterisations of the two modes of uniform convergence.

Proposition 2.8. *(i)* $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$ *if and only if for any sequence* $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$

$$\lim_{n\to\infty} d_{BL}^{\theta_n}(X_{n,\theta_n}, X_{\theta_n}) = 0.$$

(ii) $X_{n,\theta} \overset{P}{\rightrightarrows} X_\theta$ *if and only if for any sequence* $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$ *and any* $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}_{\theta_n}(\|X_{n,\theta_n} - X_{\theta_n}\| \geq \varepsilon) = 0.$$

Proof. The proof given in Kasy (2019, Lemma 1) also works in the Banachian case. □

In the remainder of this section we derive various properties of uniform convergence in probability and distribution that are analogous to the well-known properties of non-uniform convergence. In particular, we first consider a uniform version of the continuous mapping theorem which relies on stronger versions of continuity.

Proposition 2.9. *Let* $\psi : \mathcal{B} \to \tilde{\mathcal{B}}$ *where* $\tilde{\mathcal{B}}$ *is another separable Banach space.*

(i) If $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$ *and* ψ *is Lipschitz-continuous then* $\psi(X_{n,\theta}) \overset{\mathcal{D}}{\rightrightarrows} \psi(X_\theta)$.

(ii) If $X_{n,\theta} \overset{P}{\rightrightarrows} X_\theta$ *and* ψ *is uniformly continuous then* $\psi(X_{n,\theta}) \overset{P}{\rightrightarrows} \psi(X_\theta)$.

Proof. The proof in Kasy (2019, Theorem 1) also works in the Banachian case. □

In what follows we will investigate different alternative assumptions such that continuity of ψ suffices. One such assumption is tightness of the family of pushforward measures $(X_\theta(\mathbb{P}_\theta))_{\theta\in\Theta}$.

Definition 2.4. *Let* $(\mu_\theta)_{\theta\in\Theta}$ *be a family of probability measures on* $\mathcal{B}$.

(i) $(\mu_\theta)_{\theta\in\Theta}$ *is said to be* tight *if for any* $\varepsilon > 0$*, there exists a compact set* K *such that* $\sup_{\theta\in\Theta} \mu_\theta(K^c) < \varepsilon$*.* $(X_\theta)_{\theta\in\Theta}$ *is said to be* uniformly tight with respect to Θ *if the family of pushforward measures* $(X_\theta(\mathbb{P}_\theta))_{\theta\in\Theta}$ *is tight. If* Θ *is clear from the context we simply say that* $(X_\theta)_{\theta\in\Theta}$ *is uniformly tight.*

(ii) $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ *is said to be* sequentially tight with respect to Θ *if for any sequence* $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$ *the sequence of pushforward measures* $(X_{n,\theta_n}(\mathbb{P}_{\theta_n}))_{n\in\mathbb{N}}$ *is tight. If* Θ *is clear from the context we simply say that* $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ *is sequentially tight.*

(iii) $(\mu_\theta)_{\theta\in\Theta}$ *is said to be* relatively compact *if for any sequence* $(\theta_n)_{n\in\mathbb{N}}$ *there exists a subsequence* $(\theta_{k(n)})_{n\in\mathbb{N}}$*, where* $k : \mathbb{N} \to \mathbb{N}$ *is strictly increasing, such that* $\mu_{\theta_{k(n)}}$ *converges weakly to some measure* μ*, which is not necessarily in the family* $(\mu_\theta)_{\theta\in\Theta}$*.*

Prokhorov's theorem states that tightness implies relative compactness and that they are equivalent on separable and complete metric spaces; in this work, we therefore use the terms interchangeably since we only consider separable Banach and Hilbert spaces. With a uniform tightness assumption, we can perform continuous operations and preserve uniform convergence in probability just as in the non-uniform setting.

Proposition 2.10. *Let* $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ *and* $(X_\theta)_{\theta\in\Theta}$ *be random variables taking values in* $\mathcal{B}$*. Assume that* $X_{n,\theta} \overset{P}{\rightrightarrows} X_\theta$ *and* X_θ *is uniformly tight. Then, for any continuous function* $\psi : \mathcal{B} \to \tilde{\mathcal{B}}$*, where* $\tilde{\mathcal{B}}$ *is another separable Banach space, we have*

$$\psi(X_{n,\theta}) \overset{P}{\rightrightarrows} \psi(X_\theta).$$

Proof. Let $\epsilon > 0$ be given. We need to show that

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(\|\psi(X_{n,\theta}) - \psi(X_\theta)\| \geq \epsilon) \to 0$$

As X_θ is uniformly tight, for $\eta > 0$ there exists a compact set K such that

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(X_\theta \notin K_X) < \eta/2.$$

By the Heine–Cantor theorem, ψ is uniformly continuous on K_X, so there exists $\delta > 0$ such that $\|x - x'\| < \delta$ implies that $\|\psi(x) - \psi(x')\| < \epsilon$. We thus have

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(\|\psi(X_{n,\theta}) - \psi(X_\theta)\| \geq \epsilon) \leq \sup_{\theta\in\Theta} \mathbb{P}_\theta(X_\theta \notin K) + \sup_{\theta\in\Theta} \mathbb{P}_\theta(\|X_{n,\theta} - X_\theta\| \geq \delta).$$

By assumption, we can choose N sufficiently large such that for all $n \geq N$, the final term is less than $\eta/2$, resulting in the whole expression being less than η. As η was arbitrary, this proves the result. □

Bengs and Holzmann (2019) make repeated use of an alternative assumption for many of their results for real-valued random variables.

Definition 2.5. *A family of probability measures* $(\mu_\theta)_{\theta\in\Theta}$ *is* uniformly absolutely continuous with respect to the measure μ *if for any* $\varepsilon > 0$*, there exists* $\delta > 0$ *such that for any Borel set* B

$$\mu(B) < \delta \implies \sup_{\theta\in\Theta} \mu_\theta(B) < \varepsilon.$$

A family of random variables $(X_\theta)_{\theta\in\Theta}$ *is* uniformly absolutely continuous over Θ with respect to the measure μ *if the family of pushforward measures* $(X_\theta(\mathbb{P}_\theta))_{\theta\in\Theta}$ *is uniformly absolutely continuous with respect to* μ*. When* Θ *is clear from the context we simply say that* X_θ *is uniformly absolutely continuous with respect to* μ*.*

Uniform absolute continuity has previously been studied in other works such as the ones by Bogachev (2018, Section 5.6) and Doob (1994, Chapter IX, Section 4). An intuitive view of uniform absolute continuity can be given when μ is a finite measure. In this case, we can define a pseudometric d_μ on the Borel sets with $d_\mu(A, B) = \mu(A\triangle B)$, where $A\triangle B$ is the symmetric difference. Uniform absolute continuity is then uniform d_μ-continuity over θ of the collection of push-forward measures $(X_\theta(\mathbb{P}_\theta))_{\theta\in\Theta}$ viewed as mappings from the Borel sets into $\mathbb{R}$.

Another helpful perspective is in the case where for each θ, X_θ has a density f_θ with respect to a common measure μ. The following proposition shows that X_θ is uniformly absolutely continuous with respect to μ if and only if for each θ, X_θ has a density f_θ with respect to μ and the family of densities is uniformly integrable. A convenient sufficient condition for uniform integrability is the existence of $r > 0$ such that $\sup_{\theta\in\Theta} \int f_\theta^{1+r} \,\mathrm{d}\mu < \infty$.

Proposition 2.11. *If* $(X_\theta)_{\theta\in\Theta}$ *is uniformly absolutely continuous with respect to* μ*, then for each* θ X_θ *has a density* f_θ *with respect to* μ *and the family* $(f_\theta)_{\theta\in\Theta}$ *is uniformly integrable with respect to* μ*. Conversely, if for each* θ X_θ *has a density* f_θ *with respect to* μ *and the family* $(f_\theta)_{\theta\in\Theta}$ *is uniformly integrable then* $(X_\theta)_{\theta\in\Theta}$ *is uniformly absolutely continuous with respect to* μ*.*

Proof. For the first statement, note that by the Radon–Nikodym theorem, we need to show that for each θ, $\mu(B) = 0$ implies that $\mathbb{P}_\theta(X_\theta \in B) = 0$ for every Borel measurable B. This is immediate from the assumption of uniform absolute continuity (by negation) and so is the uniform integrability of the family $(f_\theta)_{\theta\in\Theta}$. The second statement follows immediately from the definitions of uniform integrability and uniform absolute continuity. □

In Bengs and Holzmann (2019) uniform absolute continuity is assumed with respect to a probability measure. For uniformly tight Banachian random variables that are uniformly absolutely continuous with respect to a σ-finite measure μ, we can show that the family is also uniformly absolutely continuous with respect to any σ-finite measure ν such that μ has a continuous density with respect to ν.

Proposition 2.12. *Assume that* $(X_\theta)_{\theta\in\Theta}$ *is uniformly absolutely continuous with respect to some* σ*-finite measure* μ*. If* ν *is another* σ*-finite measure dominating* μ *and there exists a continuous Radon-Nikodym derivative of* μ *with respect to* ν*, then* X *is uniformly absolutely continuous with respect to* ν*.*

Proof. Let $\varepsilon > 0$ be given. Because $(X_\theta)_{\theta\in\Theta}$ is uniformly tight, we can choose a compact set K, such that

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(X_\theta \notin K) < \varepsilon/2.$$

Then note that for any Borel measurable set B

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(X_\theta \in B) < \varepsilon/2 + \sup_{\theta\in\Theta} \mathbb{P}_\theta(X_\theta \in B \cap K).$$

We thus need to find δ so that $\nu(B \cap K) < \delta$ implies $\sup_{\theta\in\Theta} \mathbb{P}_\theta(X_\theta \in B \cap K) < \varepsilon/2$. Letting g denote the continuous Radon-Nikodym derivative of μ with respect to ν, we see that

$$\mu(B \cap K) = \int_{B\cap K} g \, \mathrm{d}\nu \leq \left(\sup_{x\in K} g(x)\right) \nu(B \cap K).$$

The supremum is finite by the extreme value theorem for continuous functions since K is compact. If $\sup_{x\in K} g(x) > 0$ choose δ' from the uniform absolute continuity of X with respect to μ matching $\varepsilon/2$ and set $\delta = \delta'/(\sup_{x\in K} g(x))$. Then for all B with $\nu(B) < \delta$, we have

$$\delta > \nu(B) \geq \nu(B \cap K) \geq \frac{\mu(B \cap K)}{\sup_{x\in K} g(x)} \implies \mu(B \cap K) < \delta'$$

and thus

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(X_\theta \in B \cap K) < \varepsilon/2$$

proving the result. If $\sup_{x\in K} g(x) = 0$ any δ works since $\mu(B \cap K) = 0$ implies $\sup_{\theta\in\Theta} \mathbb{P}(X_\theta \in B \cap K) = 0$. □

A consequence of the above result is that uniform absolute continuity with respect to the Lebesgue measure implies uniform absolute continuity with respect to the standard Gaussian measure. This lets us immediately apply many of the results of Bengs and Holzmann (2019) such as Theorem 4.1, when we consider a uniformly tight real-valued random variable that is uniformly absolutely continuous with respect to the Lebesgue measure.

Corollary 2.1. *A real-valued family of random variables $(X_\theta)_{\theta\in\Theta}$ is uniformly absolutely continuous with respect to the Lebesgue measure if and only if it is uniformly absolutely continuous with respect to the standard Gaussian measure.*

Proof. The statement follows immediately by the equivalence of the standard Gaussian measure and the Lebesgue measure, by the continuity of the Gaussian density and its reciprocal, and Proposition 2.12. □

We will consider sums of real-valued random variables and thus need to consider when such sums are uniformly absolutely continuous with respect to a measure. It turns out that when the random variables are independent and one of the families is uniformly absolutely continuous with respect to the Lebesgue measure, the same is true for the family of sums.

Theorem 2.9. *Let $(X_\theta)_{\theta\in\Theta}$ and $(Y_\theta)_{\theta\in\Theta}$ be two real-valued random variables such that for any $\theta \in \Theta$ X_θ and Y_θ are independent under $\mathbb{P}_\theta$. Assume that $(X_\theta)_{\theta\in\Theta}$ is uniformly absolutely continuous with respect to the Lebesgue measure. Then $(X_\theta + Y_\theta)_{\theta\in\Theta}$ is uniformly absolutely continuous with respect to the Lebesgue measure.*

Proof. Let $\varepsilon > 0$ be given and let λ denote the Lebesgue measure. We need to find $\delta > 0$ such that for any Borel measurable B with $\lambda(B) < \delta$, we have $\sup_{\theta\in\Theta} \mathbb{P}_\theta(X_\theta + Y_\theta \in B) < \varepsilon$. We can use the independence of X_θ and Y_θ to write the probability as a double-integral with respect to the pushforward measures $X_\theta(\mathbb{P}_\theta)$ and $Y_\theta(\mathbb{P}_\theta)$ as follows:

$$\mathbb{P}_\theta(X_\theta + Y_\theta \in B) = \int \mathbb{1}_B(X_\theta(\omega) + Y_\theta(\omega)) \,\mathrm{d}\mathbb{P}_\theta(\omega) = \int\int \mathbb{1}_B(x+y) \,\mathrm{d}X_\theta(\mathbb{P}_\theta)(x) \,\mathrm{d}Y_\theta(\mathbb{P}_\theta)(y).$$

Note that $\mathbb{1}_B(x+y) = \mathbb{1}_{B-y}(x)$ where $B - y := \{b - y : b \in B\}$ and that, by the translation invariance of the Lebesgue measure, $\lambda(B) = \lambda(B-y)$. As X_θ is uniformly absolutely continuous with respect to λ, there exists δ such that if $\lambda(B) < \delta$ we have

$$\begin{aligned}\sup_{\theta\in\Theta} \mathbb{P}_\theta(X_\theta + Y_\theta \in B) &\leq \sup_{\theta\in\Theta} \int \left(\sup_{\theta\in\Theta} \int \mathbb{1}_{B-y}(x) \,\mathrm{d}X_\theta(\mathbb{P}_\theta)(x) \right) \mathrm{d}Y_\theta(\mathbb{P}_\theta)(y) \\ &< \sup_{\theta\in\Theta} \int \varepsilon \,\mathrm{d}Y_\theta(\mathbb{P}_\theta)(y) < \varepsilon. \end{aligned}$$

□

Thus far, we have not discussed when we can expect uniform convergence in distribution to imply uniform convergence of distribution functions. This is exactly where we need an assumption of uniform absolute continuity. The following result is a modified version of Bengs and Holzmann (2019, Theorem 4.1), where our condition includes uniform convergence in x, rather than convergence for all x.

Proposition 2.13. *Let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ and $(X_\theta)_{\theta\in\Theta}$ be real-valued random variables. Assume that $(X_\theta)_{\theta\in\Theta}$ is uniformly absolutely continuous with respect to a continuous probability measure μ. Then $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$ if and only if*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \sup_{\theta\in\Theta} |\mathbb{P}_\theta(X_{n,\theta} \leq x) - \mathbb{P}_\theta(X_\theta \leq x)| = 0. \tag{2.30}$$

Proof. See Bengs and Holzmann (2019, Theorem 4.1) for a proof that $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$ if and only if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} |\mathbb{P}_\theta(X_{n,\theta} \leq x) - \mathbb{P}_\theta(X_\theta \leq x)| = 0$$

for all $x \in \mathbb{R}$. To show that the convergence of distribution functions is uniform, we proceed as follows. In view of the uniform absolute continuity of $(X_\theta)_{\theta\in\Theta}$ with respect to μ, for all $\varepsilon > 0$ there exists $\delta > 0$ such that for Borel measurable B with $\mu(B) < \delta$, we have $\sup_{\theta\in\Theta} \mathbb{P}_\theta(X_\theta \in B) < \varepsilon$. Let $-\infty = x_0 < x_1 < \cdots < x_m = \infty$ such that for all $i \in \{1, \ldots, m\}$, $0 < \mu((x_{i-1}, x_i]) < \delta$. We can find such a grid since μ is a continuous probability measure. For

any θ and $i \in \{1, \dots, m\}$, we thus have

$$\mathbb{P}_\theta(X_\theta \leq x_i) - \mathbb{P}_\theta(X_\theta \leq x_{i-1}) = \mathbb{P}_\theta(X_\theta \in (x_{i-1}, x_i]) < \varepsilon.$$

For $x \in (x_{i-1}, x_i]$,

$$\begin{aligned}
&\sup_{\theta\in\Theta}\{\mathbb{P}_\theta(X_{n,\theta} \leq x) - \mathbb{P}_\theta(X_\theta \leq x)\} \leq \sup_{\theta\in\Theta}\{\mathbb{P}_\theta(X_{n,\theta} \leq x_i) - \mathbb{P}_\theta(X_\theta \leq x_{i-1})\} \\
&\leq \sup_{\theta\in\Theta}\{\mathbb{P}_\theta(X_{n,\theta} \leq x_i) - \mathbb{P}_\theta(X_\theta \leq x_i)\} + \varepsilon \leq \sup_{\theta\in\Theta}|\mathbb{P}_\theta(X_{n,\theta} \leq x_i) - \mathbb{P}_\theta(X_\theta \leq x_i)| + \varepsilon,
\end{aligned}$$

and, similarly,

$$\begin{aligned}
&\sup_{\theta\in\Theta}\{\mathbb{P}_\theta(X_\theta \leq x) - \mathbb{P}_\theta(X_{n,\theta} \leq x)\} \leq \sup_{\theta\in\Theta}\{\mathbb{P}_\theta(X_\theta \leq x_i) - \mathbb{P}_\theta(X_{n,\theta} \leq x_{i-1})\} \\
&\leq \sup_{\theta\in\Theta}\{\mathbb{P}_\theta(X_\theta \leq x_{i-1}) - \mathbb{P}_\theta(X_{n,\theta} \leq x_{i-1})\} + \varepsilon \leq \sup_{\theta\in\Theta}|\mathbb{P}_\theta(X_\theta \leq x_{i-1}) - \mathbb{P}_\theta(X_{n,\theta} \leq x_{i-1})| + \varepsilon.
\end{aligned}$$

Thus,

$$\sup_{x\in\mathbb{R}}\sup_{\theta\in\Theta}|\mathbb{P}_\theta(X_{n,\theta} \leq x) - \mathbb{P}_\theta(X_\theta \leq x)| \leq \sup_{i\in\{0,\dots,m\}}\sup_{\theta\in\Theta}|\mathbb{P}_\theta(X_{n,\theta} \leq x_i) - \mathbb{P}_\theta(X_\theta \leq x_i)| + \varepsilon.$$

The first term on the right-hand side goes to 0 by assumption and ε was arbitrary, thus proving the uniform convergence. □

The final results of this section are uniform versions of Slutsky's lemma, the Weak Law of Large Numbers and the Central Limit Theorem. In the remaining results uniform tightness will play a crucial role. It is a standard result that if $(X_n)_{n\in\mathbb{N}}$ converges in distribution to X then $(X_n)_{n\in\mathbb{N}}$ is tight. We can show that analogously if $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$ and $(X_\theta)_{\theta\in\Theta}$ is uniformly tight then $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ is sequentially tight.

Proposition 2.14. *Assume that $(X_\theta)_{\theta\in\Theta}$ is uniformly tight. If $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$ then $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ is sequentially tight.*

Proof. We prove the contrapositive statement. Assume that there exists a sequence $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$ such that $(X_{n,\theta_n}(\mathbb{P}_{\theta_n}))_{n\in\mathbb{N}}$ is not tight. Let Y_n be distributed as $X_{n,\theta_n}(\mathbb{P}_{\theta_n})$ and Z_n distributed as $X(\mathbb{P}_{\theta_n})$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Since $(Y_n)_{n\in\mathbb{N}}$ is not tight, there exists a subsequence $(k(n))_{n\in\mathbb{N}}$ with $k : \mathbb{N} \to \mathbb{N}$ strictly increasing such that any further subsequence of $(Y_{k(n)})_{n\in\mathbb{N}}$ does not converge in distribution. Since $(Z_n)_{n\in\mathbb{N}}$ is tight, there exists a strictly increasing $k' : \mathbb{N} \to \mathbb{N}$ and a random variable Z such that writing $m = k \circ k'$, we have

$$d_{\mathrm{BL}}(Z_{m(n)}, Z) \to 0.$$

However, since $Y_{k(n)}$ does not have a weakly convergent subsequence, we have

$$d_{\mathrm{BL}}(Y_{m(n)}, Z) \not\to 0.$$

Thus, there exists $\varepsilon > 0$ and a strictly increasing $k'' : \mathbb{N} \to \mathbb{N}$ such that writing $l = m \circ k''$, we have for all n

$$d_{\mathrm{BL}}(Y_{l(n)}, Z) \geq \varepsilon.$$

Next choose N such that for $n \geq N$ we have

$$d_{\mathrm{BL}}(Z_{l(n)}, Z) < \varepsilon/2.$$

Then by the reverse triangle inequality

$$d_{\mathrm{BL}}(Z_{l(n)}, Y_{l(n)}) \geq \left| d_{\mathrm{BL}}(Z_{l(n)}, Z) - d_{\mathrm{BL}}(Z, Y_{l(n)}) \right| \geq \varepsilon/2$$

for all $n \geq N$. Since

$$d_{\mathrm{BL}}(Z_{l(n)}, Y_{l(n)}) = d_{\mathrm{BL}}^{\theta_{l(n)}} \left(X_{l(n),\theta_{l(n)}}, X_{\theta_{l(n)}} \right)$$

by Proposition 2.8 we cannot have $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$ proving the desired statement. □

The previous result will be required when proving the second part of the upcoming uniform version of Slutsky's lemma.

Proposition 2.15 (Uniform Slutsky's lemma)**.** *Let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$, $(Y_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ and $(X_\theta)_{\theta\in\Theta}$ be Banachian random variables. Assume that $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$ and $Y_{n,\theta} \overset{P}{\rightrightarrows} 0$. Then, the following two statements hold.*

(i) $X_{n,\theta} + Y_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$.

(ii) If $(Y_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ is a family of real-valued random variables and $(X_\theta)_{\theta\in\Theta}$ is uniformly tight, then $Y_{n,\theta}X_{n,\theta} \overset{P}{\rightrightarrows} 0$.

Proof. We first prove (i), for which we need to show that

$$\sup_{\theta\in\Theta} d_{\mathrm{BL}}^{\theta}(X_{n,\theta} + Y_{n,\theta}, X_\theta) \to 0$$

as $n \to \infty$. We have for any θ

$$d_{\mathrm{BL}}^{\theta}(X_{n,\theta} + Y_{n,\theta}, X_\theta) \leq d_{\mathrm{BL}}^{\theta}(X_{n,\theta} + Y_{n,\theta}, X_{n,\theta}) + d_{\mathrm{BL}}^{\theta}(X_{n,\theta}, X_\theta),$$

where the second term goes to 0 uniformly by assumption. It remains to show that the first term goes to 0 uniformly. Now for $f \in \mathrm{BL}_1$ we have that for any $\varepsilon > 0$ and any $x, y \in \mathcal{B}$, $\|y\| < \varepsilon$ implies $\|f(x+y) - f(x)\| \leq \varepsilon$. Hence, by using the triangle inequality for the expectation, partitioning the integral and using the uniform continuity above, we get

$$d_{\mathrm{BL}}^{\theta}(X_{n,\theta} + Y_{n,\theta}, X_{n,\theta}) \leq \varepsilon + \sup_{f\in\mathrm{BL}_1} \mathbb{E}_\theta \left| [f(X_{n,\theta} + Y_{n,\theta}) - f(X_{n,\theta})] \mathbb{1}_{\{\|Y_{n,\theta}\| > \varepsilon\}} \right|.$$

We can again apply the triangle inequality and recall that f is bounded by 1, yielding

$$\sup_{\theta\in\Theta}\sup_{f\in \mathrm{BL}_1}\mathbb{E}_\theta\left|[f(X_{n,\theta}+Y_{n,\theta})-f(X_{n,\theta})]\mathbb{1}_{\{\|Y_{n,\theta}\|>\varepsilon\}}\right| \leq 2\sup_{\theta\in\Theta}\mathbb{P}_\theta(\|Y_{n,\theta}\|>\varepsilon),$$

which goes to 0 by assumption. Since $\varepsilon > 0$ was arbitrary, we have proven the desired result.

We now turn to the proof of (ii). We will apply Proposition 2.8 and show that for any $(\theta_n)_{n\in\mathbb{N}} \subseteq \Theta$ and any $\varepsilon > 0$,

$$\mathbb{P}_{\theta_n}(\|Y_{n,\theta_n}X_{n,\theta_n}\| \geq \varepsilon) \to 0$$

as $n \to \infty$, which implies the desired result. Let $\delta > 0$ be given. By Proposition 2.14 there exists a compact set K such that

$$\sup_{n\in\mathbb{N}}\mathbb{P}_{\theta_n}(X_{n,\theta_n} \in K^c) \leq \delta/2.$$

Since K is compact, it is bounded and thus there exists $M > 0$ such that $\|x\| < M$ for all $x \in K$. By the uniform convergence in probability of Y_n to zero, we can find N such that for all $n \geq N$,

$$\mathbb{P}_{\theta_n}(|Y_{n,\theta_n}| \geq \varepsilon/M) < \delta/2.$$

Putting things together, we get, for all $n \geq N$,

$$\begin{aligned}\mathbb{P}_{\theta_n}(\|X_{n,\theta_n}Y_{n,\theta_n}\| \geq \varepsilon) &\leq \mathbb{P}_{\theta_n}(\|X_{n,\theta_n}Y_{n,\theta_n}\| \geq \varepsilon, X_{n,\theta_n} \in K) + \mathbb{P}_{\theta_n}(X_{n,\theta_n} \in K^c)\\ &\leq \mathbb{P}_{\theta_n}(|Y_{n,\theta_n}| \geq \varepsilon/M) + \sup_{n\in\mathbb{N}}\mathbb{P}_{\theta_n}(X_{n,\theta_n} \in K^c) < \delta,\end{aligned}$$

proving the result. □

We will now consider the setting of uniform convergence of averages of i.i.d. random variables, i.e. we assume that for each $\theta \in \Theta$ the sequence $(X_{n,\theta})_{n\in\mathbb{N}}$ is i.i.d. and consider the convergence of $1/n\sum_{i=1}^n X_{i,\theta}$. We first prove a small technical lemma and then apply this lemma to prove an analogue of the Law of Large numbers for uniform convergence in probability for Hilbertian random variables.

Lemma 2.8. *Let $Y_1, \ldots, Y_n$ be independent, mean zero random variables taking values in Hilbert space $\mathcal{H}$. Then*

$$\mathbb{E}\left(\left\|\sum_{i=1}^n Y_i\right\|^2\right) = \sum_{i=1}^n \mathbb{E}\|Y_i\|^2.$$

Proof. Note first that

$$\left\|\sum_{i=1}^n Y_i\right\|^2 = \sum_{i=1}^n\sum_{j=1}^n \langle Y_i, Y_j\rangle.$$

Let $(e_k)_{k\in\mathbb{N}}$ denote a basis of $\mathcal{H}$. Then for $i \neq j$

$$\mathbb{E}(\langle Y_i, Y_j\rangle) = \mathbb{E}\left(\sum_{k=1}^\infty \langle Y_i, e_k\rangle\langle Y_j, e_k\rangle\right) = \sum_{k=1}^\infty \mathbb{E}\left(\langle Y_i, e_k\rangle\langle Y_j, e_k\rangle\right) = \sum_{k=1}^\infty \mathbb{E}(\langle Y_i, e_k\rangle)\mathbb{E}(\langle Y_j, e_k\rangle)$$

but $\mathbb{E}(\langle Y_i, e_k\rangle) = 0$ for all i and k since Y_i are mean zero. □

Proposition 2.16. *Let $(X_\theta)_{\theta\in\Theta}$ be Hilbertian random variables with $\mathbb{E}_\theta(X_\theta) = 0$ for all $\theta \in \Theta$ and $\sup_{\theta\in\Theta} \mathbb{E}_\theta(\|X_\theta\|^{1+\eta}) < C$ for some $C, \eta > 0$. Let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be random variables such that for $\theta \in \Theta$ $(X_{n,\theta})_{n\in\mathbb{N}}$ is i.i.d. with the same distribution as X_θ under $\mathbb{P}_\theta$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_{i,\theta} \overset{P}{\rightrightarrows} 0.$$

Proof. We adapt the argument given in Shah and Peters (2020, Lemma 19). Defining $S_{n,\theta} := n^{-1}\sum_{i=1}^n X_{i,\theta}$, we need to show that for any $\varepsilon > 0$,

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(\|S_{n,\theta}\| \geq \varepsilon) \to 0$$

as $n \to \infty$. To this end, we let $M > 0$ and define $X_\theta^< := \mathbb{1}_{\{\|X_\theta\|\leq M\}} X_\theta$ and $X_\theta^> := \mathbb{1}_{\{\|X_\theta\|>M\}} X_\theta$ and similarly $X_{i,\theta}^<$ and $X_{i,\theta}^>$ for $i \in \mathbb{N}$. We also define $S_{n,\theta}^< := n^{-1}\sum_{i=1}^n X_{i,\theta}^<$ and $S_{n,\theta}^> := n^{-1}\sum_{i=1}^n X_{i,\theta}^>$. Note first that

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(\|X_\theta\| > M) \leq \frac{\sup_{\theta\in\Theta}\mathbb{E}_\theta\|X_\theta\|}{M} \leq \frac{C}{M},$$

hence choosing M large, we can make $\mathbb{P}_\theta(\|X_\theta\| > M)$ small uniformly in θ. Combining this with the fact that $\mathbb{E}(X_\theta^<) = -\mathbb{E}(X_\theta^>)$, we get

$$\sup_{\theta\in\Theta}\|\mathbb{E}(X_\theta^<)\| = \sup_{\theta\in\Theta}\|\mathbb{E}(X_\theta^>)\| \leq \sup_{\theta\in\Theta}\mathbb{E}\|X_\theta^>\| \leq \sup_{\theta\in\Theta}\left(\mathbb{E}\|X_\theta\|^{1+\eta}\right)^{\frac{1}{1+\eta}} \mathbb{P}_\theta(\|X_\theta\| > M)^{\frac{\eta}{1+\eta}} \leq \frac{C^2}{M}, \tag{2.31}$$

by Hölder's inequality. This implies that choosing M large we can ensure that $\sup_{\theta\in\Theta}\|\mathbb{E}(X_\theta^<)\| < \varepsilon/3$ and for these M, we have

$$\begin{aligned}\sup_{\theta\in\Theta} \mathbb{P}_\theta(\|S_{n,\theta}\| > \varepsilon) &\leq \sup_{\theta\in\Theta} \mathbb{P}_\theta(\|S_{n,\theta}^<\| > 2\varepsilon/3) + \sup_{\theta\in\Theta} \mathbb{P}_\theta(\|S_{n,\theta}^>\| > \varepsilon/3) \\ &\leq \sup_{\theta\in\Theta} \mathbb{P}_\theta(\|S_{n,\theta}^< - \mathbb{E}(X_\theta^<)\| > \varepsilon/3) + \sup_{\theta\in\Theta} \mathbb{P}_\theta(\|S_{n,\theta}^>\| > \varepsilon/3).\end{aligned}$$

By Markov's inequality and the triangle inequality

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(\|S_{n,\theta}^>\| > \varepsilon/3) \leq \frac{3\sup_{\theta\in\Theta}\mathbb{E}_\theta\|X_\theta^>\|}{\varepsilon}$$

which we have already shown in (2.31) is uniformly small when M is sufficiently large. Finally, by Markov's inequality, the triangle inequality and Lemma 2.8, we have

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(\|S_{n,\theta}^< - \mathbb{E}(X_\theta^<)\| > \varepsilon/3) \leq \frac{\sup_{\theta\in\Theta}\mathrm{E}_\theta\|S_{n,\theta}^< - \mathbb{E}(X_\theta^<)\|^2}{t^2} = \frac{\sup_{\theta\in\Theta}\mathbb{E}_\theta\|X_\theta^<\|^2}{nt^2} \leq \frac{M^2}{nt^2}$$

hence choosing n sufficiently large, we can control the final term. □

We can extend the previous result to a special class of Banach spaces under an additional tightness assumption. Recall that a Banach space $\mathcal{B}$ has a *Schauder basis* if there exists $(e_k)_{k\in\mathbb{N}}$ such that for every $v \in \mathcal{B}$ there exists a unique sequence of scalars $(\alpha_k)_{k\in\mathbb{N}}$ satisfying

$$\left\| v - \sum_{k=1}^{K} \alpha_k e_k \right\| \to 0$$

as $K \to \infty$.

Proposition 2.17. *Let $(X_\theta)_{\theta\in\Theta}$ be Banachian random variables taking values in $\mathcal{B}$ with $\mathbb{E}_\theta(X_\theta) = 0$ for all $\theta \in \Theta$ and $\sup_{\theta\in\Theta} \mathbb{E}_\theta(\|X_\theta\|^{1+\eta}) < C$ for some $C, \eta > 0$. Let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be random variables such that for $\theta \in \Theta$ $(X_{n,\theta})_{n\in\mathbb{N}}$ is i.i.d. with the same distribution as X_θ under $\mathbb{P}_\theta$. Assume further that $\mathcal{B}$ has a Schauder basis and that $(X_\theta)_{\theta\in\Theta}$ is uniformly tight. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_{i,\theta} \overset{P}{\rightrightarrows} 0.$$

Proof. For $K \in \mathbb{N}$ let P_K denote the canonical projection of $v \in \mathcal{B}$ onto the first K components of the Schauder basis, i.e. the mapping

$$v = \sum_{k=1}^{\infty} \alpha_k e_k \mapsto \sum_{k=1}^{K} \alpha_k e_k.$$

This mapping is linear and satisfies that $\sup_{K\in\mathbb{N}} \|P_K\|_{\text{op}} < \infty$ by Li and Queffélec (2017, Theorem II.2 and II.3). By the triangle inequality

$$\mathbb{P}_\theta\left(\left\|\frac{1}{n}\sum_{i=1}^{n} X_{i,\theta}\right\| \geq \varepsilon\right) \leq \mathbb{P}_\theta\left(\left\|\frac{1}{n}\sum_{i=1}^{n} P_K X_{i,\theta}\right\| \geq \varepsilon/2\right) + \mathbb{P}_\theta\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(X_{i,\theta} - P_K X_{i,\theta})\right\| \geq \varepsilon/2\right),$$

hence it is sufficient to show that the first term converges to 0 uniformly as $n \to \infty$ for fixed K and that the second term converges to 0 uniformly as $K \to \infty$. By Proposition 2.16 the first term converges to 0 for fixed K since $(P_K X_\theta)_{\theta\in\Theta}$ are concentrated on a finite-dimensional subspace of $\mathcal{B}$ and since

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta\left(\|P_K X_\theta\|^{1+\eta}\right) \leq \|P_K\|_{\text{op}}^{1+\eta} \sup_{\theta\in\Theta} \mathbb{E}_\theta\left(\|X_\theta\|^{1+\eta}\right) < \infty.$$

It remains to show that when we choose K large, the second term is small. Bogachev (2018, Theorem 2.7.10) characterises tightness of families of random variables on Banach spaces with a Schauder basis. In particular, they satisfy

$$\lim_{K\to\infty} \sup_{\theta\in\Theta} \mathbb{P}_\theta(\|X_\theta - P_K X_\theta\| > \varepsilon) = 0 \tag{2.32}$$

for every $\varepsilon > 0$. Applying Markov's inequality, partitioning the integral, applying Hölder's inequality and the triangle inequality yields that for any $t > 0$ and $\delta > 0$,

$$\begin{aligned}
&\sup_{\theta\in\Theta} \mathbb{P}_\theta \left(\left\| \frac{1}{n} \sum_{i=1}^{n} (X_{i,\theta} - P_K X_{i,\theta}) \right\| \geq t \right) \leq \frac{1}{t} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|X_\theta - P_K X_\theta\| \\
&\leq \frac{1}{t} \sup_{\theta\in\Theta} \left\{ \delta + \mathbb{E}_\theta \left(\|X_\theta - P_K X_\theta\| \mathbb{1}_{\{\|X_\theta - P_K X_\theta\| > \delta\}} \right) \right\} \\
&\leq \frac{1}{t} \sup_{\theta\in\Theta} \left\{ \delta + \left(\mathbb{E}_\theta \|X_\theta - P_K X_\theta\|^{1+\eta} \right)^{\frac{1}{1+\eta}} \left(\mathbb{P}_\theta(\|X_\theta - P_K X_\theta\| > \delta) \right)^{\frac{\eta}{1+\eta}} \right\} \\
&\leq \frac{1}{t} \left\{ \delta + (1 + \sup_{K\in\mathbb{N}} \|P_K\|_{\mathrm{op}}) C \sup_{\theta\in\Theta} \left(\mathbb{P}_\theta(\|X_\theta - P_K X_\theta\| > \delta) \right)^{\frac{\eta}{1+\eta}} \right\}.
\end{aligned}$$

By (2.32), we can choose δ and K such that the upper bound is arbitrarily small, hence we have shown the desired result. □

For the uniform central limit theorem, we only consider the Hilbertian case since this is sufficient for our needs and avoids technical problems to do with tightness and the regular (non-uniform) central limit theorem on Banach spaces. We first give some sufficient conditions for uniform convergence in distribution of Hilbertian random variables.

Proposition 2.18. *Let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ and $(X_\theta)_{\theta\in\Theta}$ be Hilbertian random variables. Assume that*

(i) for all $h \in \mathcal{H}$, $\langle X_{n,\theta}, h\rangle \overset{\mathcal{D}}{\rightrightarrows} \langle X_\theta, h\rangle$,

(ii) $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ is sequentially tight, and

(iii) $(X_\theta)_{\theta\in\Theta}$ is uniformly tight.

Then, $X_{n,\theta} \overset{\mathcal{D}}{\rightrightarrows} X_\theta$.

Proof. Let $(\theta_n)_{n\in\mathbb{N}} \subseteq \Theta$ and let Y_n have distribution $X_{n,\theta}(\mathbb{P}_{\theta_n})$ and Z_n have distribution $X_\theta(\mathbb{P}_{\theta_n})$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose for contradiction that

$$d_{\mathrm{BL}}^{\theta_n}(X_{n,\theta_n}, X_{\theta_n}) = d_{\mathrm{BL}}(Y_n, Z_n) \not\to 0$$

as $n \to \infty$. Then there exists a subsequence of Y_n and Z_n and an $\varepsilon > 0$ such that for all n

$$d_{\mathrm{BL}}(Y_{k(n)}, Z_{k(n)}) \geq \varepsilon,$$

where $k : \mathbb{N} \to \mathbb{N}$ is a strictly increasing function. By sequential tightness of $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$, there exists a subsequence of $(Y_{k(n)})_{n\in\mathbb{N}}$, represented by the index function $m = k \circ k'$ for a strictly increasing $k' : \mathbb{N} \to \mathbb{N}$ such that the subsequence $(Y_{m(n)})_{n\in\mathbb{N}}$ converges weakly to some random variable Y. By uniform tightness of X there exists a further subsequence of $(Z_{m(n)})_{n\in\mathbb{N}}$, represented by the index function $l = m \circ k''$ for a strictly increasing $k'' : \mathbb{N} \to \mathbb{N}$ such that

$(Z_{l(n)})_{n\in\mathbb{N}}$ converges weakly to some random variable Z. Note that since the range of l is a subset of the range of m, $(Y_{l(n)})_{n\in\mathbb{N}}$ also converges to Y.

We intend to show that the distributions of Z and Y are equal. The distribution of a Hilbertian random variable is completely determined by the distribution of the linear functionals (Hsing and Eubank, 2015, Theorem 7.1.2). However, for any $h \in \mathcal{H}$ and any n,

$$\begin{aligned} &d_{\mathrm{BL}}\left(\langle Y,h\rangle, \langle Z,h\rangle\right) \\ &\quad \leq d_{\mathrm{BL}}\left(\langle Y,h\rangle, \langle Y_{l(n)},h\rangle\right) + d_{\mathrm{BL}}\left(\langle Y_{l(n)},h\rangle, \langle Z_{l(n)},h\rangle\right) + d_{\mathrm{BL}}\left(\langle Z_{l(n)},h\rangle, \langle Z,h\rangle\right). \end{aligned}$$

The first and third term of the right-hand side go to zero by definition and the middle term goes to zero by assumption (i). Now,

$$d_{\mathrm{BL}}(Y_{l(n)}, Z_{l(n)}) \leq d_{\mathrm{BL}}(Y_{l(n)}, Z) + d_{\mathrm{BL}}(Z, Z_{l(n)}).$$

Hence, we can choose N making $l(N)$ large enough that the RHS is smaller than $\varepsilon/2$. This is a contradiction since we chose k such that $d_{\mathrm{BL}}(Y_{k(n)}, Z_{k(n)}) \geq \varepsilon$ for all $n \in \mathbb{N}$ but $(l(n))_{n\in\mathbb{N}} \subseteq (k(n))_{n\in\mathbb{N}}$. □

We can now prove a uniform central limit theorem in Hilbert spaces.

Proposition 2.19. *Let $(X_\theta)_{\theta\in\Theta}$ be Hilbertian random variables with $\mathbb{E}_\theta(X_\theta) = 0$ for all θ and $\sup_{\theta\in\Theta} \mathbb{E}_\theta(\|X_\theta\|^{2+\eta}) < K$ for some $K, \eta > 0$. Denote $(\mathscr{C}_\theta)_{\theta\in\Theta}$ the family of covariance operators of X_θ under each $\mathbb{P}_\theta$, i.e. $\mathscr{C}_\theta = \mathbb{E}_\theta(X_\theta \otimes X_\theta)$. Let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be random variables such that for $\theta \in \Theta$ $(X_{n,\theta})_{n\in\mathbb{N}}$ is i.i.d. with the same distribution as X_θ under $\mathbb{P}_\theta$. Assume further that for some orthonormal basis $(e_k)_{k=1}^\infty$ of $\mathcal{H}$*

$$\lim_{K\to\infty} \sup_{\theta\in\Theta} \sum_{k=K}^{\infty} \langle \mathscr{C}_\theta e_k, e_k\rangle = 0. \tag{2.33}$$

Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{i,\theta} \overset{\mathcal{D}}{\rightrightarrows} Z$$

where the distribution of Z under $\mathbb{P}_\theta$ is $\mathcal{N}(0, \mathscr{C}_\theta)$.

Proof. We intend to apply Proposition 2.18 and thus check the conditions. For the first condition, let $h \in \mathcal{H}$ be given and let $Y_n = \langle X_n, h\rangle$ and let Y be distributed as $\langle \mathcal{N}(0, \mathscr{C}_\theta), h\rangle$ under $\mathbb{P}_\theta$, i.e. as $\mathcal{N}(0, \langle \mathscr{C}_\theta h, h\rangle)$. Note that

$$\left\langle \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{i,\theta}, h \right\rangle = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_{i,\theta}$$

hence by Proposition 2.8 it is sufficient for the first condition that for any $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$

$$\lim_{n\to\infty} d_{\mathrm{BL}}^{\theta_n}(Y_n, Y) = 0.$$

Suppose for contradiction that there exists a sequence $(\theta_n)_{n\in\mathbb{N}}$ such that the limit does not equal 0. Then there exists an $\varepsilon > 0$ and a strictly increasing function $m : \mathbb{N} \to \mathbb{N}$ such that

$$d_{\mathrm{BL}}^{\theta_{m(n)}}(Y_n, Y) \geq \varepsilon$$

for any $n \in \mathbb{N}$. Denoting for $n \in \mathbb{N}$, $\sigma^2_{\theta_{m(n)}} := \langle \mathscr{C}_{\theta_{m(n)}} h, h\rangle$, we note that the sequence $\left(\sigma^2_{\theta_{m(n)}}\right)_{n\in\mathbb{N}}$ is bounded by assumption and hence by the Bolzano–Weierstrass theorem it has a convergent subsequence, i.e. there exists $\sigma^2 \geq 0$ and a strictly increasing $m' : \mathbb{N} \to \mathbb{N}$ such that letting $l = m' \circ m$, $\sigma^2_{\theta_{l(n)}} \to \sigma^2$. Letting W denote a random variable with distribution $\mathcal{N}(0, \sigma^2)$ for any $\mathbb{P}_\theta$, by Scheffé's lemma this implies that

$$\lim_{n\to\infty} d_{\mathrm{BL}}^{\theta_{l(n)}}(Y, W) = 0.$$

Further, the Lindeberg–Feller theorem (Durrett, 2019, Theorem 3.4.10) yields that

$$\lim_{n\to\infty} d_{\mathrm{BL}}^{\theta_{l(n)}}(Y_n, W) = 0,$$

since Lyapunov's condition is fulfilled by the uniform bound on the $(2+\eta)$th moment of X_θ. Because the range of l is contained in the range of m, this is a contradiction, hence the first condition is fulfilled.

The third condition follows immediately from the assumption in (2.33) by Bogachev (2018, Proposition 2.5.2, Lemma 2.7.20). Define $S_{n,\theta} := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_{i,\theta}$ for $n \in \mathbb{N}$. The second condition follows by the same assumption and theorems by observing that $\mathbb{E}_\theta \|S_{n,\theta}\|^2$ is bounded by the same constant bounding $\mathbb{E}_\theta \|X_\theta\|^2$ and that

$$\mathrm{Cov}_\theta(S_{n,\theta}) = \frac{1}{n} \sum_{i=1}^n \mathrm{Cov}_\theta(X_{i,\theta}) = \mathscr{C}_\theta.$$

This shows that the family of measures $(X_{n,\theta}(\mathbb{P}_\theta))_{n\in\mathbb{N},\theta\in\Theta}$ is tight which implies the second condition. □

2.9 Proofs of results in Sections 2.3.2 and 2.4

This section contains the proofs of all results in Sections 2.3.2 and 2.4 except Proposition 2.1 which is proven in Section 2.7.2. The proofs are self-contained, but readers new to the field may find the following references helpful. For general results about random variables on metric spaces (Slutsky's theorem, etc.) see Billingsley (1999, Chapter 1). For more specific results about Hilbertian random variables, Bochner integrals and operators on Hilbert spaces, see Hsing and Eubank (2015, Chapter 2, 4, 7). For existence and construction of conditional expectations on Hilbert spaces, see Scalora (1961, Chapter 2). In this section, we sometimes omit the subscript P when it is clear from the context.

2.9.1 Derivation of (2.7)

We first prove a small lemma.

Lemma 2.9. *Let $x_1, \ldots, x_n$ be elements of a Hilbert space $\mathcal{H}$. Then*

$$\left\| \sum_{i=1}^{n} x_i \right\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle x_i, x_j \rangle$$

and the non-zero eigenvalues of the operator

$$\mathscr{A} := \sum_{i=1}^{n} x_i \otimes x_i$$

equal the eigenvalues of the matrix A with entries

$$A_{ij} := \langle x_i, x_j \rangle.$$

Proof. The first claim is immediate, since

$$\left\| \sum_{i=1}^{n} x_i \right\|^2 = \left\langle \sum_{i=1}^{n} x_i, \sum_{j=1}^{n} x_j \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle x_i, x_j \rangle.$$

For the second claim, note that we can write the operator $\mathscr{A}$ as $\mathscr{B}^*\mathscr{B}$ where $\mathscr{B} : \mathcal{H} \to \mathbb{R}^n$ is an operator given by

$$\mathscr{B}h = \begin{pmatrix} \langle x_1, h \rangle \\ \vdots \\ \langle x_n, h \rangle \end{pmatrix}$$

with adjoint $\mathscr{B}^*$ given by

$$\mathscr{B}^* v = \sum_{i=1}^{n} v_i x_i.$$

The result now follows since $A = \mathscr{B}\mathscr{B}^*$. □

Applying the first result of the lemma to the sequence $1/\sqrt{n}\mathscr{R}_i$ for $i = 1, \ldots, n$ viewed as Hilbert–Schmidt operators from $\mathcal{H}_X$ to $\mathcal{H}_Y$, we get that

$$T_n = \frac{1}{n} \left\| \sum_{i=1}^{n} \mathscr{R}_i \right\|^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \mathscr{R}_i, \mathscr{R}_j \rangle = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \hat{\varepsilon}_i, \hat{\varepsilon}_j \rangle \langle \hat{\xi}_i, \hat{\xi}_j \rangle.$$

Applying the second result of the lemma to the sequence $1/\sqrt{n-1}\mathscr{R}_i - \bar{\mathscr{R}}$, we get that the eigenvalues of

$$\hat{\mathscr{C}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathscr{R}_i - \bar{\mathscr{R}}) \otimes_{\mathrm{HS}} (\mathscr{R}_i - \bar{\mathscr{R}})$$

equal the eigenvalues of the matrix A with entries

$$A_{ij} := \frac{1}{n-1}\langle \mathscr{R}_i - \bar{\mathscr{R}}, \mathscr{R}_j - \bar{\mathscr{R}} \rangle$$

Using bilinearity of the inner product, we can expand and see that

$$A = \Gamma - J\Gamma - \Gamma J + J\Gamma J$$

as desired.

2.9.2 Derivation of (2.11)

Proof. Fix $n \geq 2$ and write $p := 1 + a(n)$. Let $(\tilde{x}_i, \tilde{y}_i, \tilde{z}_i)_{i=1}^n$ denote mean-centred observations, so e.g. $\tilde{z}_i = z_i - \sum_{j=1}^n z_j/n$, and let $\tilde{X}^{(n)} = (\tilde{x}_1, \ldots, \tilde{x}_n)^\top \in \mathbb{R}^n$. Let $W_n \in \mathbb{R}^{n \times p}$ be the design matrix with ith row given by $(\tilde{y}_i, \tilde{z}_{i1}, \ldots, \tilde{z}_{ia(n)})$, and let $\hat{\theta}_n \in \mathbb{R}$ be the first component of the coefficient vector from regressing $\tilde{X}^{(n)}$ onto W_n, so $\hat{\theta}_n := \{(W_n^\top W_n)^{-1} W_n^\top \tilde{X}^{(n)}\}_1$. Further, let $P_n \in \mathbb{R}^{n \times n}$ be the orthogonal projection onto the column space of W_n. Then

$$\psi_n^{\mathrm{OLS}} = \mathbb{1}_{\{|\hat{\theta}_n| \geq t_{n-p-1}(\alpha/2)\hat{\sigma}_{W,n}\|(I-P_n)\tilde{X}^{(n)}\|_2/\sqrt{n-p-1}\}},$$

where $t_{n-p}(\alpha/2)$ is the upper $\alpha/2$-point of a t distribution on $n-p$ degrees of freedom, and $\hat{\sigma}^2_{W,n} := \{(W_n^\top W_n)^{-1}\}_{11}$. Fix $Q \in \mathcal{Q}$; in the following we will suppress dependence on this for notational simplicity. Then there exists $r \in \mathbb{N}$ such that

$$\theta := \frac{\mathrm{Cov}(Y, X \mid Z)}{\mathrm{Var}(X \mid Z)} = \frac{\mathrm{Cov}(Y, X \mid Z_1, \ldots, Z_r)}{\mathrm{Var}(X \mid Z_1, \ldots, Z_r)} > 0,$$

and so for n such that $a(n) > r$, $\hat{\theta}_n \,|\, W_n \sim \mathcal{N}(\theta, \sigma^2 \hat{\sigma}^2_{W,n})$ where

$$\sigma^2 := \mathrm{Var}(X \mid Y, Z) = \mathrm{Var}(X \mid Y, Z_1, \ldots, Z_r) > 0.$$

Note that $\|(I - P_n)\tilde{X}^{(n)}\|_2^2/\sigma^2 \sim \chi^2_{n-p-1}$, and so by the weak law of large numbers and the continuous mapping theorem, $\|(I-P_n)\tilde{X}^{(n)}\|_2/\sqrt{n-p-1} \xrightarrow{P} \sigma$. To show that $\mathbb{P}(\psi_n^{\mathrm{OLS}} = 1) \to 1$, it therefore suffices to show that $\hat{\sigma}^2_{W,n} \xrightarrow{P} 0$.

Now writing $\Sigma_n = \mathrm{Cov}(Y, Z_1, \ldots, Z_{a(n)})$, we have that $W_n^\top W_n$ has a Wishart distribution on $n-1$ degrees of freedom: $W_n^\top W_n \sim W_p(\Sigma_n, n-1)$. Thus, $(\Sigma_n^{-1})_{11}/\hat{\sigma}^2_{W,n} \sim \chi^2_{n-p}$ and $(\Sigma_n^{-1})_{11} = \mathrm{Var}(Y \mid Z_1, \ldots, Z_r) = \mathrm{Var}(Y \mid Z) < \infty$. We therefore see that as $n \to \infty$ and hence $n - p \to \infty$, we have $\hat{\sigma}^2_{W,n} \xrightarrow{P} 0$ as required. □

2.9.3 Proofs of results in Section 2.4.2

In this section we provide proofs of Theorems 2.3 and 2.2. The proofs rely heavily on the theory developed in Section 2.8.

Auxiliary lemmas

We first prove some auxiliary lemmas that will be needed for the upcoming proofs.

Lemma 2.10. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables defined on $(\Omega, \mathcal{F})$ equipped with a family of probability measures $(\mathbb{P}_\theta)_{\theta\in\Theta}$. Let X be another real-valued random variable on the same space and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a sequence of sub-σ-algebras of $\mathcal{F}$. If $\mathbb{E}_\theta(|X_n|\,|\,\mathcal{F}_n) \overset{P}{\rightrightarrows} 0$ then $X_n \overset{P}{\rightrightarrows} 0$.*

Proof. Let $\epsilon > 0$ be given. By Markov's inequality

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(|X_n| \geq \epsilon) \leq \sup_{\theta\in\Theta} \mathbb{P}_\theta(|X_n| \wedge \epsilon \geq \epsilon) \leq \sup_{\theta\in\Theta} \frac{\mathbb{E}_\theta(|X_n| \wedge \epsilon)}{\epsilon}.$$

We will be done if we can show that $\sup_{\theta\in\Theta} \mathbb{E}_\theta(|X_n| \wedge \epsilon) \to 0$ as $n \to \infty$. Note that by monotonicity of conditional expectations, for each $\theta \in \Theta$ we have

$$\mathbb{E}_\theta(|X_n| \wedge \epsilon\,|\,\mathcal{F}_n) \leq \mathbb{E}_\theta(\epsilon\,|\,\mathcal{F}_n) = \epsilon,$$

and

$$\mathbb{E}_\theta(|X_n| \wedge \epsilon\,|\,\mathcal{F}_n) \leq \mathbb{E}_\theta(|X_n|\,|\,\mathcal{F}_n).$$

Combining both of the above expressions, we get

$$\mathbb{E}_\theta(|X_n| \wedge \epsilon\,|\,\mathcal{F}_n) \leq \mathbb{E}_\theta(|X_n|\,|\,\mathcal{F}_n) \wedge \epsilon.$$

This lets us write by the tower property and monotonicity of integrals,

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta(|X_n| \wedge \epsilon) = \sup_{\theta\in\Theta} \mathbb{E}_\theta[\mathbb{E}_\theta(|X_n| \wedge \epsilon\,|\,\mathcal{F}_n)] \leq \sup_{\theta\in\Theta} \mathbb{E}_\theta[\mathbb{E}_\theta(|X_n|\,|\,\mathcal{F}_n) \wedge \epsilon].$$

Let $Y_n := \mathbb{E}_\theta(|X_n|\,|\,\mathcal{F}_n) \wedge \epsilon$ and let $\delta > 0$ be given. Then

$$\begin{aligned}\sup_{\theta\in\Theta} \mathbb{E}_\theta(Y_n) &\leq \sup_{\theta\in\Theta} \mathbb{E}_\theta\left(Y_n \mathbb{1}_{\{Y_n < \delta/2\}}\right) + \sup_{\theta\in\Theta} \mathbb{E}_\theta\left(Y_n \mathbb{1}_{\{Y_n \geq \delta/2\}}\right) \\ &\leq \frac{\delta}{2} + \epsilon \sup_{\theta\in\Theta} \mathbb{P}_\theta(Y_n \geq \delta/2).\end{aligned}$$

By assumption, for any $\eta > 0$, we can choose $N \in \mathbb{N}$ so that for all $n \geq N$, we can make $\sup_{\theta\in\Theta} \mathbb{E}_\theta(|X_n|\,|\,\mathcal{F}_n) \geq \delta/2) < \eta$. Thus, choosing N to parry $\eta = \frac{\epsilon}{2\epsilon}$, we get

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta(|Y_n|) < \delta,$$

proving the desired result. □

Lemma 2.11. *Let X and Y be random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a Hilbert space $\mathcal{H}$. Let $\mathcal{D}$ be a sub-σ-algebra of $\mathcal{F}$ so that X is $\mathcal{D}$-measurable. Assume*

that $\mathbb{E}(\|X\|)$, $\mathbb{E}(\|Y\|)$ *and* $\mathbb{E}(\|X\|\|Y\|)$ *all exist. Then*

$$\mathbb{E}(\langle X, Y\rangle \,|\, \mathcal{D}) = \langle X, \mathbb{E}(Y \,|\, \mathcal{D})\rangle.$$

Proof. To show the result, we need to show that $\langle X, \mathbb{E}(Y \,|\, \mathcal{D})\rangle$ is $\mathcal{D}$-measurable and that integrals over $\mathcal{D}$-sets of $\langle X, Y\rangle$ and $\langle X, \mathbb{E}(Y \,|\, \mathcal{D})\rangle$ coincide. $\langle X, \mathbb{E}(Y \,|\, \mathcal{D})\rangle$ is $\mathcal{D}$-measurable by continuity of the inner product and the fact that X and $\mathbb{E}(Y \,|\, \mathcal{D})$ are $\mathcal{D}$-measurable by assumption and definition, respectively. By expanding the inner product in an orthonormal basis $(e_k)_{k\in\mathbb{N}}$ of $\mathcal{H}$, we get

$$\begin{aligned}
&\int_D \langle X, Y\rangle \,\mathrm{d}\mathbb{P} = \int_D \sum_{k=1}^{\infty} \langle X, e_k\rangle\langle Y, e_k\rangle \,\mathrm{d}\mathbb{P} = \sum_{k=1}^{\infty} \int_D \mathbb{E}(\langle X, e_k\rangle\langle Y, e_k\rangle \,|\, \mathcal{D}) \,\mathrm{d}\mathbb{P} \\
&= \sum_{k=1}^{\infty} \int_D \langle X, e_k\rangle\langle \mathbb{E}(Y \,|\, \mathcal{D}), e_k\rangle \,\mathrm{d}\mathbb{P} = \int_D \sum_{k=1}^{\infty} \langle X, e_k\rangle\langle \mathbb{E}(Y \,|\, \mathcal{D}), e_k\rangle \,\mathrm{d}\mathbb{P} = \int_D \langle X, \mathbb{E}(Y \,|\, \mathcal{D})\rangle \,\mathrm{d}\mathbb{P},
\end{aligned}$$

by using the interchangeability of sums and integrals and the property

$$\mathbb{E}(\langle Y, e_i\rangle \,|\, \mathcal{D}) = \langle \mathbb{E}(Y \,|\, \mathcal{D}), e_i\rangle$$

of conditional expectations on Hilbert spaces.

□

Lemma 2.12. *Let* q *denote the function that maps a self-adjoint, positive semidefinite, trace-class operator,* $\mathscr{C}$ *on a separable Hilbert space* $\mathcal{H}$*, to the* $1-\alpha$ *quantile of the* $\|\mathcal{N}(0, \mathscr{C})\|^2$ *distribution. Then* q *is continuous in trace norm and the restriction of* q *to a bounded subset* $\mathcal{C}$ *of covariance operators satisfying*

$$\lim_{N\to\infty} \sup_{\mathscr{C}\in\mathcal{C}} \sum_{k=K}^{\infty} \langle \mathscr{C} e_k, e_k\rangle = 0 \tag{2.34}$$

for some orthonormal basis $(e_k)_{k=1}^{\infty}$ *of* $\mathcal{H}$*, is uniformly continuous in trace norm.*

Proof. Let $\mathscr{C}_n$ be a sequence of self-adjoint, positive semidefinite, trace-class operators converging to $\mathscr{C}$ in trace norm. Then by Bogachev (2018, Theorem 2.7.21) $\mathcal{N}(0, \mathscr{C}_n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathscr{C})$ and by the continuous mapping theorem we have $\|\mathcal{N}(0, \mathscr{C}_n)\|^2 \xrightarrow{\mathcal{D}} \|\mathcal{N}(0, \mathscr{C})\|^2$. This implies the convergence of the quantile functions by the Portmanteau theorem and Vaart (1998, Lemma 21.2) and hence q is continuous.

By the Heine–Cantor theorem, the restriction of q to the closure of $\mathcal{C}$ is uniformly continuous if $\mathcal{C}$ is relatively compact. Restricting q further to $\mathcal{C}$ preserves the uniform continuity. Bogachev (2018, Proposition 2.5.2) states that equation (2.34) exactly characterises the relatively compact sets of trace class operators. □

Lemma 2.13. *Let* $\Theta \subseteq \mathbb{R}_+$ *and let* $(\mu_\theta)_{\theta\in\Theta}$ *be the family of probability distributions on* $\mathbb{R}$ *where for each* $\theta \in \Theta$*,* μ_θ *denotes the distribution of* θZ *where* $Z \sim \chi^2_1$*. If* Θ *is bounded away from* 0*, the family is uniformly absolutely continuous with respect to the Lebesgue measure* λ*.*

Proof. Note that the density f_θ of μ_θ with respect to the Lebesgue measure is

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\theta}} \frac{e^{-\frac{1}{2\theta}x}}{\sqrt{x}}.$$

We will apply Proposition 2.11 by showing that $\sup_{\theta\in\Theta} \int f_\theta^{3/2} \, \mathrm{d}\lambda < \infty$, which is sufficient for uniform integrability by Bogachev (2007, Example 4.5.10). We see that

$$\int f_\theta(x)^{3/2} \, \mathrm{d}\lambda = \frac{1}{\sqrt[4]{6\pi^3\theta^3}} \int \frac{e^{-\frac{3}{4\theta}x}}{\sqrt[4]{x^3}} \, \mathrm{d}\lambda,$$

and we recognise the final integral as the unnormalised density of a $\Gamma(1/4, 3/(4\theta))$ random variable. Thus,

$$\int f_\theta(x)^{3/2} \, \mathrm{d}\lambda = \frac{1}{\sqrt[4]{6\pi^3\theta^3}} \frac{\Gamma(1/4)\sqrt[4]{4\theta}}{\sqrt[4]{3}} = \frac{\Gamma(1/4)}{\sqrt[4]{6\pi^3\theta^2}}.$$

This is finite for all $\theta \in \Theta$ since Θ is bounded away from zero, proving the desired result. □

Lemma 2.14. *Let X be a uniformly tight with respect to index family Θ (see Definition 2.4), real-valued and non-negative random variable that is uniformly absolutely continuous with respect to the Lebesgue measure. Then so is $\sqrt{X}$.*

Proof. Let $\epsilon > 0$ be given and let λ denote the Lebesgue measure. We need to find $\delta > 0$ such that for any Borel measurable B,

$$\lambda(B) < \delta \implies \sup_{\theta\in\Theta} \mathbb{P}_\theta(\sqrt{X} \in B) < \epsilon.$$

For each measurable B, we define $B^2 := \{b^2 \ : \ b \in \mathbb{R}\}$. Then $\mathbb{P}_\theta(\sqrt{X} \in B) = \mathbb{P}_\theta(X \in B^2)$ and by the uniform tightness of X, we can find $M > 0$ such that

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(X \in B^2) \leq \sup_{\theta\in\Theta} \mathbb{P}_\theta(X \in B^2 \cap [0, M]) + \epsilon/2.$$

By the uniform absolute continuity of X with respect to λ, we can find δ' such that $\lambda(B) < \delta'$ implies $\sup_{\theta\in\Theta} \mathbb{P}_\theta(X \in B) < \epsilon/2$. Note that for any such B, by the regularity of the Lebesgue measure, we can find an open set $U \supseteq B$ such that $\lambda(U \setminus B) < \delta' - \lambda(B)$. This implies that $\lambda(U) < \delta'$. For every open U, by Carothers (2000, Theorem 4.6), we can find a countable union of disjoint open intervals $(I_j)_{j=1}^\infty$, where $I_j = (a_j, b_j)$, such that $U = \bigcup_{j=1}^\infty I_j$. Note that U^2 also covers B^2 since if $x \in U$, x is in at least one of the intervals I_j, and thus x^2 is in I_j^2. Combining these observations, we get that

$$\begin{aligned}\lambda(B^2 \cap [0, M]) \leq \lambda(U^2 \cap [0, M]) &= \sum_{j=1}^\infty \lambda(I_j^2 \cap [0, M]) = \sum_{j=1}^\infty (\min(M, b_j^2) - a_j^2) \\ &= \sum_{j=1}^\infty (\min(\sqrt{M}, b_j) + a_j)(\min(\sqrt{M}, b_j) - a_j) \leq 2\sqrt{M} \sum_{j=1}^\infty b_j - a_j < 2\sqrt{M}\delta'.\end{aligned}$$

Thus letting $\delta = \delta'/(2\sqrt{M})$, we see that for all B with $\lambda(B) < \delta$, we also have $\lambda(B^2 \cap [0, M]) < \delta'$, and hence

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(X \in B^2 \cap [0, M]) < \epsilon/2,$$

proving the statement. □

Lemma 2.15. *Let $(X_\theta)_{\theta\in\Theta}$ be Hilbertian random variables with values in $\mathcal{H}$. Assume that for every $\theta \in \Theta$, $\mathbb{E}_\theta(X_\theta) = 0$, $\sup_{\theta\in\Theta} \mathbb{E}\|X_\theta\|^2 < \infty$ and that there exists a basis $(e_k)_{k\in\mathbb{N}}$ of $\mathcal{H}$ such that*

$$\lim_{K\to\infty} \sup_{\theta\in\Theta} \sum_{k=K}^{\infty} \mathbb{E}(\langle X_\theta, e_k\rangle^2) = 0.$$

Then the family $(X_\theta \otimes X_\theta)_{\theta\in\Theta}$ is uniformly tight when viewed as random variables in the Banach space of trace-class operators on $\mathcal{H}$.

Proof. By Fugarolas and Cobos (1983, Proposition 3.1) $(e_k \otimes e_j)_{(k,j)\in\mathbb{N}^2}$ is a Schauder basis for the Banach space of trace-class operators on $\mathcal{H}$. Thus, Bogachev (2018, Theorem 2.7.10) yields that we need to show that

$$\lim_{r\to\infty} \sup_{\theta\in\Theta} \mathbb{P}_\theta(\|X_\theta \otimes X_\theta\|_{\mathrm{TR}} > r) = 0$$

and for all $\varepsilon > 0$

$$\lim_{K\to\infty} \sup_{\theta\in\Theta} \mathbb{P}_\theta(\|X_\theta \otimes X_\theta - P_K(X_\theta \otimes X_\theta)\|_{\mathrm{TR}} > \varepsilon) = 0,$$

where P_K denotes the projection onto the K first basis vectors in the space of trace-class operators. An application of Markov's inequality yields immediately that

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(\|X_\theta \otimes X_\theta\|_{\mathrm{TR}} > r) \leq \frac{\sup_{\theta\in\Theta} \mathbb{E}_\theta\|X_\theta\|^2}{r}$$

hence the first condition is satisfied by the assumed uniform upper bound on $\mathbb{E}_\theta\|X_\theta\|^2$. For $K = m^2$, $m \in \mathbb{N}$, note that

$$\begin{aligned}\|X_\theta \otimes X_\theta - P_K(X_\theta \otimes X_\theta)\|_{\mathrm{TR}} &= \left\| \left(\sum_{j=m}^{\infty} \langle X_\theta, e_j\rangle e_j \right) \otimes \left(\sum_{k=m}^{\infty} \langle X_\theta, e_k\rangle e_k \right) \right\|_{\mathrm{TR}} \\ &= \left\| \sum_{j=m}^{\infty} \langle X_\theta, e_j\rangle e_j \right\|^2 = \sum_{j=m}^{\infty} \langle X_\theta, e_j\rangle^2,\end{aligned}$$

where the final equality is by Parseval's identity. Using this, the second condition is satisfied by assumption, since, by Markov's inequality, for all $\epsilon > 0$,

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta \left(\sum_{j=m}^{\infty} \langle X_\theta, e_j\rangle^2 > \epsilon \right) \leq \frac{\sup_{\theta\in\Theta} \sum_{j=m}^{\infty} \mathbb{E}_\theta \left(\langle X_\theta, e_j\rangle^2 \right)}{\epsilon}.$$

□

Proof of Theorem 2.2

Proof. Throughout the proof we omit the subscript P from ε, ξ, f, g.

Convergence of $\mathscr{T}_n$. We have that

$$\mathscr{T}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathscr{R}_i = \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\otimes\xi_i}_{=:U_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(z_i)-\hat{f}(z_i))\otimes(g(z_i)-\hat{g}(z_i))}_{=:a_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(z_i)-\hat{f}(z_i))\otimes\xi_i}_{=:b_n} + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\varepsilon_i\otimes g(z_i)-\hat{g}(z_i))}_{=:c_n}.$$

Since

$$\mathbb{E}(\varepsilon_i\otimes\xi_i) = \mathbb{E}((X-\mathbb{E}(X\mid Z))\otimes(Y-\mathbb{E}(Y\mid Z))) = \mathbb{E}(\mathrm{Cov}(X,Y\mid Z)) = 0$$

because $X \perp\!\!\!\perp Y \mid Z$, Proposition 2.19 yields that U_n converges uniformly in distribution to the desired Gaussian over $\tilde{\mathcal{P}}_0$. By Proposition 2.15, if a_n, b_n and c_n all converge to 0 uniformly in probability, we will have shown the desired result. We establish this by looking at the Hilbert–Schmidt norm of the sequences, since uniform convergence of the norms to 0 implies uniform convergence of the sequences to 0. For a_n, using properties of the Hilbert–Schmidt norm and the Cauchy–Schwarz inequality yields

$$\begin{aligned}
\|a_n\|_{\mathrm{HS}} &= \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(z_i)-\hat{f}(z_i))\otimes(g(z_i)-\hat{g}(z_i))\right\|_{\mathrm{HS}} \\
&\le \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\|(f(z_i)-\hat{f}(z_i))\otimes(g(z_i)-\hat{g}(z_i))\|_{\mathrm{HS}} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\|(f(z_i)-\hat{f}(z_i))\|\|g(z_i)-\hat{g}(z_i)\| \\
&\le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|(f(z_i)-\hat{f}(z_i))\|^2\sum_{i=1}^{n}\|g(z_i)-\hat{g}(z_i)\|^2} = \sqrt{nM_{n,P}^{f}M_{n,P}^{g}}.
\end{aligned}$$

By assumption $nM_{n,P}^{f}M_{n,P}^{g} \overset{P}{\rightrightarrows} 0$ and Proposition 2.10 yields that the same is true for $\sqrt{nM_{n,P}^{f}M_{n,P}^{g}}$. This implies that $\|a_n\|_{\mathrm{HS}} \overset{P}{\rightrightarrows} 0$ as desired.

To establish that $\|b_n\|_{\mathrm{HS}} \overset{P}{\rightrightarrows} 0$, we will instead show that the square of the Hilbert–Schmidt norm goes to 0. This implies that $\|b_n\|_{\mathrm{HS}} \overset{P}{\rightrightarrows} 0$ by the same arguments about $x \mapsto \sqrt{x}$ as above. We will show that $\mathbb{E}_P(\|b_n\|_{\mathrm{HS}}^2 \mid X^{(n)}, Z^{(n)}) \overset{P}{\rightrightarrows} 0$, where $X^{(n)} = (x_1,\ldots,x_n)$ and $Z^{(n)} = (x_1,\ldots,x_n)$, which then implies the desired result by Lemma 2.10. For every $P \in \tilde{\mathcal{P}}_0$

we have

$$\begin{aligned}
\mathbb{E}_P(\|b_n\|_{\mathrm{HS}}^2 \,|\, X^{(n)}, Z^{(n)}) &= \frac{1}{n}\mathbb{E}_P\left(\left\|\sum_{i=1}^n (f(z_i) - \hat{f}(z_i)) \otimes \xi\right\|_{\mathrm{HS}}^2 \,|\, X^{(n)}, Z^{(n)}\right) \\
&= \frac{1}{n}\sum_{j=1}^n\sum_{i=1}^n \mathbb{E}_P\left(\langle (f(z_i) - \hat{f}(z_i)) \otimes \xi_i, (f(z_j) - \hat{f}(z_j)) \otimes \xi_j\rangle_{\mathrm{HS}} \,|\, X^{(n)}, Z^{(n)}\right) \\
&= \frac{1}{n}\sum_{j=1}^n\sum_{i=1}^n \mathbb{E}_P\left(\langle f(z_i) - \hat{f}(z_i), f(z_j) - \hat{f}(z_j)\rangle\langle \xi_i, \xi_j\rangle \,|\, X^{(n)}, Z^{(n)}\right) \\
&= \frac{1}{n}\sum_{j=1}^n\sum_{i=1}^n \langle f(z_i) - \hat{f}(z_i), f(z_j) - \hat{f}(z_j)\rangle \mathbb{E}_P\left(\langle \xi_i, \xi_j\rangle \,|\, X^{(n)}, Z^{(n)}\right), \qquad (2.35)
\end{aligned}$$

where the penultimate equality uses the fact that $\langle x_1 \otimes y_1, x_2 \otimes y_2\rangle_{\mathrm{HS}} = \langle x_1, x_2\rangle\langle y_1, y_2\rangle$. The final equality holds since the terms involving $f(z_i) - \hat{f}(z_i)$ are measurable with respect to the σ-algebra generated by $X^{(n)}$ and $Z^{(n)}$. The term $\langle \xi_i, \xi_j\rangle$ only depends on Z_i and Z_j of the conditioning variables, so we can omit the remaining variables from the conditioning expression. Recall that $\xi_i = Y_i - \mathbb{E}_P(Y_i \,|\, Z_i)$. For $i \neq j$, by using that $\mathbb{E}_P(Y_i \,|\, Z_i) = \mathbb{E}_P(Y_i \,|\, Z_i, Z_j)$ since Z_j is independent of (Y_i, Z_i) and Lemma 2.11, we get

$$\begin{aligned}
\mathbb{E}_P[\langle \xi_i, \xi_j\rangle \,|\, X^{(n)}, Z^{(n)}] &= \mathbb{E}_P[\langle Y_i, Y_j\rangle - \langle Y_i, \mathbb{E}_P(Y_j \,|\, Z_j)\rangle - \langle \mathbb{E}_P(Y_i \,|\, Z_i), Y_j\rangle \\
&\qquad + \langle \mathbb{E}_P(Y_i \,|\, Z_i), \mathbb{E}_P(Y_j \,|\, Z_j)\rangle \,|\, Z_i, Z_j] \\
&= \mathbb{E}_P(\langle Y_i, Y_j\rangle \,|\, Z_i, Z_j) - \langle \mathbb{E}_P(Y_i \,|\, Z_i, Z_j), \mathbb{E}(Y_j \,|\, Z_i, Z_j)\rangle.
\end{aligned}$$

We will show that this is zero. By assumption $(Y_i, Z_i) \perp\!\!\!\perp (Y_j, Z_j)$, so applying the usual laws of conditional independence, we get $Y_i \perp\!\!\!\perp Y_j \,|\, (Z_i, Z_j)$. Take now some orthonormal basis for $\mathcal{H}_Y$, $(e_k)_{k\in\mathbb{N}}$, and expand $\langle Y_i, Y_j\rangle$ to get

$$\mathbb{E}_P(\langle Y_i, Y_j\rangle \,|\, Z_i, Z_j) = \mathbb{E}_P\left(\sum_{k=1}^\infty \langle Y_i, e_k\rangle\langle Y_j, e_k\rangle \,|\, Z_i, Z_j\right) = \sum_{k=1}^\infty \mathbb{E}_P\left(\langle Y_i, e_k\rangle\langle Y_j, e_k\rangle \,|\, Z_i, Z_j\right).$$

For all k, $\langle Y_i, e_k\rangle \perp\!\!\!\perp \langle Y_j, e_k\rangle \,|\, (Z_i, Z_j)$, so $\mathbb{E}(\langle Y_i, e_k\rangle\langle Y_j, e_k\rangle \,|\, Z_i, Z_j)$ factorises, and we get

$$\begin{aligned}
&\sum_{k=1}^\infty \mathbb{E}_P\left(\langle Y_i, e_k\rangle\langle Y_j, e_k\rangle \,|\, Z_i, Z_j\right) = \sum_{k=1}^\infty \mathbb{E}_P(\langle Y_i, e_k\rangle \,|\, Z_i, Z_j)\mathbb{E}_P(\langle Y_j, e_k\rangle \,|\, Z_i, Z_j) \\
&= \sum_{k=1}^\infty \langle \mathbb{E}_P(Y_i \,|\, Z_i, Z_j), e_k\rangle\langle \mathbb{E}_P(Y_j \,|\, Z_i, Z_j), e_k\rangle = \langle \mathbb{E}_P(Y_i \,|\, Z_i, Z_j), \mathbb{E}(Y_j \,|\, Z_i, Z_j)\rangle,
\end{aligned}$$

where we have used that $\mathbb{E}_P(\langle Y, e_k\rangle \,|\, Z_i, Z_j) = \langle \mathbb{E}_P(Y \,|\, Z_i, Z_j), e_k\rangle$ by Lemma 2.11. We can thus omit all terms from the sum in (2.35) where $i \neq j$ and get

$$\mathbb{E}_P(\|b_n\|_{\mathrm{HS}}^2 \,|\, X^{(n)}, Z^{(n)}) = \frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}^{(n)}(z_i)\|_X^2 \mathbb{E}_P\left(\|\xi_i\|_Y^2 \,|\, Z_i\right) = \tilde{M}_n^f \overset{P}{\rightrightarrows} 0,$$

by assumption. An analogous argument can be repeated for c_n, thus proving the desired result.

Convergence of $\hat{\mathscr{C}}$. For simplicity, we prove convergence where $\hat{\mathscr{C}}$ is instead defined as the estimate where we divide by n instead of $n-1$ since this does not affect the asymptotics.

By the above and Proposition 2.15, since $(\mathcal{N}(0, \mathscr{C}_P))_{P \in \tilde{\mathcal{P}}_0}$ is uniformly tight by Bogachev (2018, Proposition 2.5.2, Lemma 2.7.20), we have

$$\frac{1}{n}\sum_{i=1}^{n} \mathscr{R}_i = \frac{1}{\sqrt{n}} \mathscr{T}_n \overset{P}{\rightrightarrows} 0.$$

By Proposition 2.10, this implies that the second term in the definition of $\hat{\mathscr{C}}$ converges to 0 uniformly in probability since the mapping $(\mathscr{A}, \mathscr{B}) \mapsto \mathscr{A} \otimes_{\text{HS}} \mathscr{B}$ is continuous. It remains to show that the first term in the definition of $\hat{\mathscr{C}}$ converges to $\mathscr{C}$. The proof is similar to the proof of Theorem 6 in (Shah and Peters, 2020) and relies on expanding the first term $\frac{1}{n}\sum_{i=1}^{n} \mathscr{R}_i \otimes_{\text{HS}} \mathscr{R}_i$ to yield

$$\frac{1}{n}\sum_{i=1}^{n}[(f(z_i) - \hat{f}(z_i)) \otimes (g(z_i) - \hat{g}(z_i)) + (f(z_i) - \hat{f}(z_i)) \otimes \xi_i + \varepsilon_i \otimes (g(z_i) - \hat{g}(z_i)) + \varepsilon_i \otimes \xi_i]^{\otimes_{\text{HS}}2},$$

where $\mathscr{A}^{\otimes_{\text{HS}}2} = \mathscr{A} \otimes_{\text{HS}} \mathscr{A}$. Expanding this even further yields 16 terms of which 15 go to zero. The non-zero term is

$$\text{I}_n = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i \otimes \xi_i)^{\otimes_{\text{HS}}2} \overset{P}{\rightrightarrows} \mathbb{E}_P\left((\varepsilon_i \otimes \xi_i)^{\otimes_{\text{HS}}2}\right) = \mathscr{C},$$

by Proposition 2.17 and Lemma 2.15 and the assumed tightness condition. For the remaining 15 terms, we will argue by taking trace norms and applying the triangle inequality to reduce the number of cases. This leaves us with 8 terms and 5 cases (by symmetry of f and ε, g and ξ) that we need to argue converge to 0 uniformly in probability.

The first case is

$$\begin{aligned}
\text{II}_n &= \left\| \frac{1}{n}\sum_{i=1}^{n}[(f(z_i) - \hat{f}(z_i)) \otimes (g(z_i) - \hat{g}(z_i))]^{\otimes_{\text{HS}}2} \right\|_{\text{TR}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left\| [(f(z_i) - \hat{f}(z_i)) \otimes (g(z_i) - \hat{g}(z_i))]^{\otimes_{\text{HS}}2} \right\|_{\text{TR}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\| f(z_i) - \hat{f}(z_i) \otimes g(z_i) - \hat{g}(z_i) \right\|_{\text{HS}}^2 = \frac{1}{n}\sum_{i=1}^{n} \|f(z_i) - \hat{f}(z_i)\|^2 \|g(z_i) - \hat{g}(z_i)\|^2 \\
&\leq n M_{n,P}^f M_{n,P}^g \overset{P}{\rightrightarrows} 0,
\end{aligned}$$

where the final inequality uses that for positive sequences $\sum a_n b_n \leq \sum a_n \sum b_n$, which can be seen by noting that every term on the left-hand side also appears on the right-hand side. For

the second case we have, by applying the Cauchy–Schwarz inequality,

$$\begin{aligned}\mathrm{III}_n &= \left\| \frac{1}{n}\sum_{i=1}^n [(f(z_i) - \hat{f}(z_i)) \otimes \xi_i] \otimes_{\mathrm{HS}} [(g(z_i) - \hat{g}(z_i)) \otimes \varepsilon_i] \right\|_{\mathrm{TR}} \\ &\leq \frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\| \|g(z_i) - \hat{g}(z_i)\| \|\varepsilon_i\| \|\xi_i\| \\ &\leq \sqrt{\underbrace{\left(\frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\|^2 \|g(z_i) - \hat{g}(z_i)\|^2\right)}_{=:\tilde{a}_n} \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \|\varepsilon_i\|^2 \|\xi_i\|^2\right)}_{=:\tilde{U}_n}}.\end{aligned}$$

By Cauchy–Schwarz, we have $\tilde{a}_n \leq n M_{n,P}^f M_{n,P}^g \overset{P}{\rightrightarrows} 0$. We have $\tilde{U}_n \overset{P}{\rightrightarrows} \|\mathscr{C}\|_{\mathrm{TR}}$ by Proposition 2.16. The family $(\|\mathscr{C}\|_{\mathrm{TR}})_{P \in \tilde{\mathcal{P}}_0}$ is uniformly tight by the assumption that $\mathbb{E}(\|\varepsilon_P\|^{2+\eta} \|\xi_P\|^{2+\eta})$ is uniformly bounded, since this also yields a bound on $\mathbb{E}(\|\varepsilon_P\|^2 \|\xi_P\|^2) = \|\mathscr{C}\|_{\mathrm{TR}}$ thus Proposition 2.10 yields that $\sqrt{\tilde{a}_n \tilde{U}_n} \overset{P}{\rightrightarrows} 0$.

The remaining three cases have an f and a g variant where the roles of f and g and ε and ξ are swapped. We only show one variant of each, since the arguments are identical. The f-variant of the third case is

$$\mathrm{IV}_n^f = \left\| \frac{1}{n}\sum_{i=1}^n [(f(z_i) - \hat{f}(z_i)) \otimes \xi_i]^{\otimes_{\mathrm{HS}} 2} \right\|_{\mathrm{TR}} \leq \frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\|^2 \|\xi_i\|^2 =: \tilde{b}_n.$$

If we can show that $\mathbb{E}(\tilde{b}_n \mid X^{(n)}, Z^{(n)}) \overset{P}{\rightrightarrows} 0$, we have that $\tilde{b}_n \overset{P}{\rightrightarrows} 0$ by Lemma 2.10 and hence $\mathrm{IV}_n^f \overset{P}{\rightrightarrows} 0$. This holds since

$$\mathbb{E}_P(\tilde{b}_n \mid X^{(n)}, Z^{(n)}) = \frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\|^2 \mathbb{E}_P\left(\|\xi_i\|^2 \mid X^{(n)}, Z^{(n)}\right) = \tilde{M}_{n,P}^f \overset{P}{\rightrightarrows} 0,$$

by assumption.

The f-variant of the fourth case is, by applying the Cauchy–Schwarz inequality,

$$\begin{aligned}\mathrm{V}_n^f &= \left\| \frac{1}{n}\sum_{i=1}^n [(f(z_i) - \hat{f}(z_i)) \otimes (g(z_i) - \hat{g}(z_i))] \otimes_{\mathrm{HS}} [(f(z_i) - \hat{f}(z_i)) \otimes \xi_i] \right\|_{\mathrm{TR}} \\ &\leq \frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\|^2 \|g(z_i) - \hat{g}(z_i)\| \|\xi_i\| \\ &\leq \sqrt{\underbrace{\left(\frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\|^2 \|g(z_i) - \hat{g}(z_i)\|^2\right)}_{\tilde{a}_n} \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\|^2 \|\xi_i\|^2\right)}_{\tilde{b}_n}}.\end{aligned}$$

We saw above that $\tilde{a}_n \overset{P}{\rightrightarrows} 0$ and $\tilde{b}_n \overset{P}{\rightrightarrows} 0$, hence by Proposition 2.10, $\sqrt{\tilde{a}_n \tilde{b}_n} \overset{P}{\rightrightarrows} 0$.

For the f-variant of the fifth and final case, we get, by applying the Cauchy–Schwarz inequality again,

$$\begin{aligned}\mathrm{VI}_n^f &= \left\| \frac{1}{n}\sum_{i=1}^n [(f(z_i) - \hat{f}(z_i)) \otimes \xi_i] \otimes_{\mathrm{HS}} [\varepsilon_i \otimes \xi_i] \right\|_{\mathrm{TR}} \leq \frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\| \|\varepsilon_i\| \|\xi_i\|^2 \\ &\leq \sqrt{\underbrace{\left(\frac{1}{n}\sum_{i=1}^n \|f(z_i) - \hat{f}(z_i)\|^2 \|\xi_i\|^2\right)}_{\tilde{b}_n} \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \|\varepsilon_i\|^2 \|\xi_i\|^2\right)}_{\tilde{U}_n}}.\end{aligned}$$

We can repeat the arguments used above yielding $\sqrt{\tilde{a}_n \tilde{U}_n} \overset{P}{\rightrightarrows} 0$ to show that $\sqrt{\tilde{b}_n \tilde{U}_n} \overset{P}{\rightrightarrows} 0$ hence $\mathrm{VI}_n^f \overset{P}{\rightrightarrows} 0$ as desired. □

Proof of Theorem 2.3

Proof. Let W be distributed as $\|\mathcal{N}(0, \mathscr{C}_P)\|_{\mathrm{HS}}^2$ when the background measure is $\mathbb{P}_P$. Recalling the notation from Lemma 2.12, since

$$\mathbb{P}_P(\psi_n = 1) = \mathbb{P}_P(T_n > q(\hat{\mathscr{C}}))$$

we need to show that

$$\lim_{n\to\infty} \sup_{P \in \tilde{\mathcal{P}}_0} \left| \mathbb{P}_P(T_n > q(\hat{\mathscr{C}})) - \alpha \right| = 0,$$

which amounts to finding, for each $\epsilon > 0$, an $N \in \mathbb{N}$, such that for all $n \geq N$,

$$\sup_{P \in \tilde{\mathcal{P}}_0} \mathbb{P}_P(T_n > q(\hat{\mathscr{C}})) < \alpha + \epsilon \tag{2.36}$$

and

$$\inf_{P \in \tilde{\mathcal{P}}_0} \mathbb{P}_P(T_n > q(\hat{\mathscr{C}})) > \alpha - \epsilon. \tag{2.37}$$

To show (2.36), take $\delta > 0$ (to be fixed later). If $|q(\hat{\mathscr{C}}) - q(\mathscr{C}_P)| < \delta$ and $T_n > q(\hat{\mathscr{C}})$, then $T_n > q(\mathscr{C}_P) - \delta$, so

$$\mathbb{P}_P(T_n > q(\hat{\mathscr{C}})) \leq \mathbb{P}_P(T_n > q(\mathscr{C}_P) - \delta) + \mathbb{P}_P(|q(\hat{\mathscr{C}}) - q(\mathscr{C}_P)| \geq \delta).$$

Taking suprema and rewriting, we get

$$\begin{aligned}\sup_{P \in \tilde{\mathcal{P}}_0} \mathbb{P}_P(T_n > q(\hat{\mathscr{C}}_P)) \leq{}& \overbrace{\sup_{P \in \tilde{\mathcal{P}}_0} [\mathbb{P}_P(T_n > q(\mathscr{C}_P) - \delta) - \mathbb{P}_P(W > q(\mathscr{C}_P) - \delta)]}^{=:\mathrm{I}_n} \\ &+ \underbrace{\sup_{P \in \tilde{\mathcal{P}}_0} [\mathbb{P}_P(W > q(\mathscr{C}_P) - \delta) - \alpha]}_{=:\mathrm{II}_n} + \underbrace{\sup_{P \in \tilde{\mathcal{P}}_0} \mathbb{P}_P(|q(\hat{\mathscr{C}}) - q(\mathscr{C}_P)| \geq \delta)}_{=:\mathrm{III}_n} + \alpha.\end{aligned}$$

We seek to show that, if n is sufficiently large, we can make each of the terms I_n, II_n and III_n less than $\epsilon/3$ such that

$$\sup_{P\in\tilde{\mathcal{P}}_0} \mathbb{P}_P(T_n > q(\hat{\mathscr{C}})) < \alpha + \epsilon,$$

as desired.

We note first that

$$\begin{aligned} |\mathrm{I}_n| &\leq \sup_{P\in\tilde{\mathcal{P}}_0} |\mathbb{P}_P(T_n^{1/2} > \{q(\mathscr{C}_P) - \delta\}^{1/2}) - \mathbb{P}_P(W^{1/2} > \{q(\mathscr{C}_P) - \delta\}^{1/2})| \\ &\leq \sup_{P\in\tilde{\mathcal{P}}_0} \sup_{x\in\mathbb{R}} |\mathbb{P}_P(T_n^{1/2} > x) - \mathbb{P}_P(W^{1/2} > x)|. \end{aligned} \tag{2.38}$$

For each $P \in \tilde{\mathcal{P}}_0$, W has the same distribution as

$$\sum_{k=1}^{\infty} \lambda_k^P V_k^2,$$

where λ_k^P is the kth eigenvalue of $\mathscr{C}_P$ and $(V_k)_{k\in\mathbb{N}}$ is a sequence of independent standard Gaussian random variables. We have assumed that the operator norm of $(\mathscr{C}_P)_{P\in\tilde{\mathcal{P}}_0}$ is bounded away from zero which implies that λ_1^P is bounded away from zero. Thus, the family $(\lambda_1^P V_1^2)_{P\in\tilde{\mathcal{P}}_0}$ is uniformly absolutely continuous with respect to the Lebesgue measure by Lemma 2.13. Theorem 2.9 yields that W is also uniformly absolutely continuous with respect to the Lebesgue measure and Lemma 2.14 yields that the same is true for $W^{1/2}$, since W is uniformly tight by the assumed uniform bound on $\mathbb{E}_P(\|\epsilon_P\|^2\|\xi_P\|^2)$. Further, Corollary 2.1 yields that $W^{1/2}$ is also uniformly absolutely continuous with respect to the standard Gaussian on $\mathbb{R}$. Proposition 2.9 and Theorem 2.2 $T_n^{1/2} \overset{\mathcal{D}}{\rightrightarrows} W^{1/2}$ since $\|\cdot\|_{\mathrm{HS}}$ is Lipschitz. Finally, since we argued that $W^{1/2}$ is uniformly absolutely continuous with respect to the standard Gaussian on $\mathbb{R}$, Proposition 2.13 yields that we can make the bound in (2.38) less than $\epsilon/3$ for n sufficiently large.

For the II_n term, recall that $\alpha = \mathbb{P}_P(W > q(\mathscr{C}_P))$, and thus

$$\mathbb{P}_P(W > q(\mathscr{C}_P) - \delta) - \alpha = \mathbb{P}_P(W \in [q(\mathscr{C}_P) - \delta, q(\mathscr{C}_P)]).$$

By the uniform absolute continuity of W with respect to the Lebesgue measure λ, we may fix δ such that $\sup_{P\in\tilde{\mathcal{P}}_0} \mathbb{P}_P(W \in B) < \epsilon/3$ whenever $\lambda(B) < 2\delta$. This implies that $\mathrm{II}_n < \epsilon/3$.

For the III_n term, Theorem 2.2 yields $\hat{\mathscr{C}} \overset{P}{\rightrightarrows} \mathscr{C}_P$ and since Lemma 2.12 yields that q is uniformly continuous, Proposition 2.9 yields $q(\hat{\mathscr{C}}) \overset{P}{\rightrightarrows} q(\mathscr{C}_P)$. Thus, the third term is less than $\epsilon/3$ when n is large enough.

To show (2.37), note first that, as before, if $|q(\hat{\mathscr{C}}) - q(\mathscr{C}_P)| < \delta$ and $T_n > q(\mathscr{C}_P) + \delta$, then $T_n > q(\hat{\mathscr{C}})$ and hence

$$\begin{aligned} \mathbb{P}_P(T_n > q(\hat{\mathscr{C}})) &\geq \mathbb{P}_P((T_n > q(\mathscr{C}_P) + \delta) \cap (|q(\hat{\mathscr{C}}) - q(\mathscr{C}_P)| < \delta)) \\ &\geq \mathbb{P}_P(T_n > q(\mathscr{C}_P) + \delta) - \mathbb{P}_P(|q(\hat{\mathscr{C}}) - q(\mathscr{C}_P)| \geq \delta). \end{aligned} \tag{2.39}$$

The final step uses that for any measurable sets A and B,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) = \mathbb{P}(A) - \mathbb{P}(B^c) + 1 - \mathbb{P}(A \cup B) \geq \mathbb{P}(A) - \mathbb{P}(B^c).$$

This lets us continue using similar arguments as for (2.36), proving the statement. □

2.9.4 Proof of Theorem 2.4

Proof. To argue that the modified GHCM satisfies (2.15), we can repeat the arguments of Theorem 2.2 and Theorem 2.3 replacing conditioning on $X^{(n)}$ and $Z^{(n)}$ with conditioning on $Z^{(n)}$ and A and conditioning on $Y^{(n)}$ and $Z^{(n)}$ with conditioning on $Z^{(n)}$ and A.

For the first claim that $\tilde{\mathscr{T}}_n \overset{\mathcal{D}}{\rightrightarrows} \mathcal{N}(0, \mathscr{C}_P)$, we can repeat the decomposition of the proof of Theorem 2.2 and write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathscr{R}_i - \mathscr{K}_P) = \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\varepsilon_i \otimes \xi_i - \mathscr{K}_P)}_{=:U_n} + a_n + b_n + c_n,$$

where a_n, b_n and c_n are as in the proof of Theorem 2.2. We have $U_n \overset{\mathcal{D}}{\rightrightarrows} \mathcal{N}(0, \mathscr{C}_P)$ over $\mathcal{Q}$ by Proposition 2.19 $a_n \overset{P}{\rightrightarrows} 0$ over $\mathcal{Q}$ by the same argument as in the proof of Theorem 2.2. The argument of the proof of Theorem 2.2 to show that $b_n \overset{P}{\rightrightarrows} 0$ and $c_n \overset{P}{\rightrightarrows} 0$ will also work here if we replace conditioning as we did for the first claim.

For the second claim that $\|\hat{\mathscr{C}} - \mathscr{C}\|_{\mathrm{TR}} \overset{P}{\rightrightarrows} 0$, note that by the $\tilde{\mathscr{T}}_n$ result, Proposition 2.15 and Proposition 2.10,

$$\frac{1}{n} \sum_{i=1}^{n} \mathscr{R}_i = \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathscr{R}_i - \mathscr{K}_P) + \mathscr{K}_P \overset{P}{\rightrightarrows}_{\mathcal{Q}} \mathscr{K}_P.$$

Hence, by Proposition 2.10,

$$\left(\frac{1}{n} \sum_{i=1}^{n} \mathscr{R}_i \right) \otimes_{\mathrm{HS}} \left(\frac{1}{n} \sum_{i=1}^{n} \mathscr{R}_i \right) \overset{P}{\rightrightarrows}_{\mathcal{Q}} \mathscr{K}_P \otimes_{\mathrm{HS}} \mathscr{K}_P,$$

since the mapping $(\mathscr{A}, \mathscr{B}) \mapsto \mathscr{A} \otimes_{\mathrm{HS}} \mathscr{B}$ is continuous. We can now repeat the remaining arguments of the proof of Theorem 2.2 while again replacing conditioning as we did in the proof of the first claim to yield the desired result.

For the final claim that for large enough n the GHCM has power greater than β over alternatives where $\|\sqrt{n}\mathscr{K}_P\|_{\mathrm{HS}} > c$, let W be distributed as $\|\mathcal{N}(0, \mathscr{C}_P)\|_{\mathrm{HS}}^2$ when the background measure is $\mathbb{P}_P$ for $P \in \mathcal{Q}$. Let q denote the mapping that sends a covariance operator $\mathscr{C}$ to the $1 - \alpha$ quantile of the distribution of $\|\mathcal{N}(0, \mathscr{C})\|_{\mathrm{HS}}^2$ as in Lemma 2.12. By similar arguments as (2.39) in the proof of Theorem 2.3, we get that for any $\delta > 0$, $c > 0$ and $n \in \mathbb{N}$,

$$\inf_{P \in \mathcal{Q}_{c,n}} \mathbb{P}_P(T_n > q(\hat{\mathscr{C}})) \geq \inf_{P \in \mathcal{Q}_{c,n}} \mathbb{P}_P(T_n > q(\mathscr{C}_P) + \delta) - \sup_{P \in \mathcal{Q}_{c,n}} \mathbb{P}_P(|q(\hat{\mathscr{C}}) - q(\mathscr{C}_P)| \geq \delta).$$

Defining $\tilde{T}_n^{1/2} := \|\tilde{\mathscr{T}}_n\|_{\mathrm{HS}}$, by the reverse triangle inequality

$$T_n^{1/2} = \left\|\tilde{\mathscr{T}}_n + \sqrt{n}\mathscr{K}_P\right\|_{\mathrm{HS}} \geq \left|\tilde{T}_n^{1/2} - \sqrt{n}\|\mathscr{K}_P\|_{\mathrm{HS}}\right| \geq \sqrt{n}\|\mathscr{K}_P\|_{\mathrm{HS}} - \tilde{T}_n^{1/2},$$

and hence

$$\inf_{P\in\mathcal{Q}_{c,n}} \mathbb{P}_P(T_n > q(\mathscr{C}_P) + \delta) \geq \inf_{P\in\mathcal{Q}_{c,n}} \mathbb{P}_P(\sqrt{n}\|\mathscr{K}_P\|_{\mathrm{HS}} - \tilde{T}_n^{1/2} > \{q(\mathscr{C}_P) + \delta\}^{1/2}).$$

Now since we are taking an infimum over a set where $\sqrt{n}\|\mathscr{K}_P\|_{\mathrm{HS}} > c$, we have

$$\inf_{P\in\mathcal{Q}_{c,n}} \mathbb{P}_P(\sqrt{n}\|\mathscr{K}_P\|_{\mathrm{HS}} - \tilde{T}_n^{1/2} > \{q(\mathscr{C}_P) + \delta\}^{1/2}) \geq \inf_{P\in\mathcal{Q}_{c,n}} \mathbb{P}_P(c - \tilde{T}_n^{1/2} > \{q(\mathscr{C}_P) + \delta\}^{1/2}),$$

and thus combining all the above yields

$$\begin{aligned}
&\inf_{P\in\mathcal{Q}_{c,n}} \mathbb{P}_P(T_n > q(\hat{\mathscr{C}}_P)) \\
&\geq \overbrace{\inf_{P\in\mathcal{Q}_{c,n}} [\mathbb{P}_P(c - \tilde{T}_n^{1/2} > \{q(\mathscr{C}_P) + \delta\}^{1/2}) - \mathbb{P}_P(c - W^{1/2} > \{q(\mathscr{C}_P) + \delta\}^{1/2})]}^{=:\mathrm{I}_n} \\
&+ \underbrace{\inf_{P\in\mathcal{Q}_{c,n}} \mathbb{P}_P(c - W^{1/2} > \{q(\mathscr{C}_P) + \delta\}^2)}_{=:\mathrm{II}_n} - \underbrace{\sup_{P\in\mathcal{Q}_{c,n}} \mathbb{P}_P(|q(\hat{\mathscr{C}}) - q(\mathscr{C}_P)| \geq \delta)}_{=:\mathrm{III}_n}.
\end{aligned}$$

If we can show that for n sufficiently large we can make $\mathrm{I}_n + \mathrm{II}_n + \mathrm{III}_n \geq \beta$, we will be done.

For the I_n term, we can write

$$\mathrm{I}_n \geq -\sup_{P\in\mathcal{Q}_{c,n}} \sup_{x\in\mathbb{R}} |\mathbb{P}_P(\tilde{T}_n^{1/2} < x) - \mathbb{P}_P(W^{1/2} < x)|.$$

By the first claim proven above and Proposition 2.9, $\tilde{T}_n^{1/2} \overset{\mathcal{D}}{\rightrightarrows} W^{1/2}$. We can therefore repeat the arguments used to deal with the I_n term in the proof of Theorem 2.3 to see that for n sufficiently large we have $\mathrm{I}_n \geq -(1-\beta)/3$.

For the II_n term, we can write

$$\mathrm{II}_n = 1 - \sup_{P\in\mathcal{Q}_{c,n}} \mathbb{P}_P(W^{1/2} + \{q(\mathscr{C}_P) + \delta\}^{1/2} \geq c).$$

Hence, by uniform tightness of $(W^{1/2} + \{q(\mathscr{C}_P) + \delta\}^{1/2})_{P\in\mathcal{Q}}$ we can find c such that

$$\sup_{P\in\mathcal{Q}_{c,n}} \mathbb{P}_P(W^{1/2} + \{q(\mathscr{C}_P) + \delta\}^{1/2} \geq c) < (1-\beta)/3$$

which implies $\mathrm{II}_n > 1 - (1-\beta)/3$.

For the III_n term, we can repeat the arguments for the III_n term in the proof of Theorem 2.3 to show that $\mathrm{III}_n \overset{P}{\rightrightarrows} 0$. Hence, for sufficiently large n, we have $\mathrm{III}_n > -(1-\beta)/3$.

Putting things together, we have for n sufficiently large that

$$\inf_{P \in \mathcal{Q}_{c,n}} \mathbb{P}_P(T_n > q(\hat{\mathscr{C}})) \geq \beta. \qquad \square$$

2.9.5 Proof of Theorem 2.5 and related results

We first prove a representer theorem (Kimeldorf and Wahba, 1970; Schölkopf et al., 2001) for scalar-on-function regression which we use to provide bounds on the in-sample error of the Hilbertian linear model in Lemma 2.17.

Lemma 2.16. *Let $\mathcal{H}$ denote a Hilbert space with norm $\|\cdot\|$, $x_1, \ldots, x_n \in \mathbb{R}$, $z_1, \ldots, z_n \in \mathcal{H}$ and $\gamma > 0$ Let K be an $n \times n$ matrix where $K_{i,j} := \langle z_i, z_j \rangle$ and let $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$. Then $\hat{\beta}$ minimises*

$$L_1(\beta) = \sum_{i=1}^{n} (x_i - \langle \beta, z_i \rangle)^2 + \gamma \|\beta\|^2$$

over $\beta \in \mathcal{H}$ if and only if $\hat{\beta} = \sum_{i=1}^n \hat{\alpha}_i z_i$ and $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)^\top \in \mathbb{R}^n$ minimises

$$L_2(\alpha) = \|x - K\alpha\|_2^2 + \gamma \alpha^\top K \alpha$$

over $\mathbb{R}^n$ where $\|\cdot\|_2$ denotes the standard Euclidean norm on $\mathbb{R}^n$.

Proof. Assume that $\hat{\beta}$ minimises L_1. Write $\hat{\beta} = u + v$ where $u \in \mathcal{U} := \operatorname{span}(z_1, \ldots, z_n)$ and $v \in \mathcal{U}^\perp$. Since

$$\langle \hat{\beta}, z_i \rangle = \langle u, z_i \rangle,$$

the first term of L_1 only depends on the quantity u. Also, by Pythagoras' theorem,

$$\|\hat{\beta}\|^2 = \|u\|^2 + \|v\|^2 \geq \|u\|^2.$$

Thus, $v = 0$ by optimality of $\hat{\beta}$, and so $\hat{\beta}$ can be written

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i z_i$$

for some $\hat{\alpha} \in \mathbb{R}^n$. But now that $\hat{\beta}$ is known to have this form, it can be seen that $\hat{\alpha}^\top K \hat{\alpha} = \|\hat{\beta}\|^2$ and

$$\sum_{i=1}^{n} (x_i - \langle \hat{\beta}, z_i \rangle)^2 = \sum_{i=1}^{n} \left(x_i - \sum_{j=1}^{n} \hat{\alpha}_j \langle z_j, z_i \rangle \right)^2 = \|x - K\hat{\alpha}\|_2^2,$$

hence $\hat{\alpha}$ minimises L_2.

Assume now that $\hat{\alpha} \in \mathbb{R}^n$ minimises L_2 and $\hat{\beta} = \sum_{i=1}^n \hat{\alpha}_i z_i$. Clearly, $L_2(\hat{\alpha}) = L_1(\hat{\beta})$. For any $\tilde{\beta} \in \mathcal{H}$, we can write $\tilde{\beta} = \tilde{u} + \tilde{v}$ with $\tilde{u} \in \mathcal{U}$ and $\tilde{v} \in \mathcal{U}^\perp$ as before. By similar arguments as above,

$$L_1(\tilde{\beta}) \geq L_1(\tilde{u}).$$

However, $\tilde{u} = \sum_{i=1}^n \tilde{\alpha}_i z_i$, hence by optimality of $\hat{\alpha}$, we have

$$L_1(\tilde{u}) = L_2(\tilde{\alpha}) \geq L_2(\hat{\alpha}) = L_1(\hat{\beta}),$$

proving that $\hat{\beta}$ minimises L_1 as desired. □

Lemma 2.17. *Let $n \in \mathbb{N}$ be fixed. Consider the estimator $\hat{\mathscr{S}}$ (2.19) in the Hilbertian linear model which is a function of $x_1, \ldots, x_n, z_1, \ldots, z_n$ and let $\sigma^2 > 0$ be such that $\mathbb{E}(\|\varepsilon\|^2 \mid Z) \leq \sigma^2$ almost surely. Let K be the $n \times n$ matrix where $K_{ij} := \langle z_i, z_j \rangle$ and let $(\hat{\mu}_i)_{i=1}^n$ denote the eigenvalues of K. Then, letting $Z^{(n)} := (z_1, \ldots, z_n)$,*

$$\frac{1}{n}\mathbb{E}\left(\sum_{i=1}^n \|\mathscr{S}(z_i) - \hat{\mathscr{S}}(z_i)\|^2 \mid Z^{(n)}\right) \leq \frac{\sigma^2}{\gamma}\frac{1}{n}\sum_{i=1}^n \min(\hat{\mu}_i/4, \gamma) + \|\mathscr{S}\|_{HS}^2 \frac{\gamma}{4n} \tag{2.40}$$

almost surely.

Proof. Let $(e_k)_{k\in\mathbb{N}}$ denote a basis of $\mathcal{H}_X$ and write $\langle \cdot, \cdot \rangle_X$ and $\langle \cdot, \cdot \rangle_Z$ for the inner products and $\|\cdot\|_X$ and $\|\cdot\|_Z$ for the norms on $\mathcal{H}_X$ and $\mathcal{H}_Z$, respectively. Then

$$\begin{aligned}\sum_{i=1}^n \|\mathscr{S}(z_i) - \hat{\mathscr{S}}(z_i)\|_X^2 &= \sum_{k=1}^\infty \sum_{i=1}^n (\langle \mathscr{S}(z_i), e_k \rangle_X - \langle \hat{\mathscr{S}}(z_i), e_k \rangle_X)^2 \\ &= \sum_{k=1}^\infty \sum_{i=1}^n (\langle z_i, \mathscr{S}^*(e_k) \rangle_Z - \langle z_i, \hat{\mathscr{S}}^*(e_k) \rangle_Z)^2 \end{aligned} \tag{2.41}$$

and similarly we can rewrite the penalised square-error criterion in (2.19) as

$$\sum_{i=1}^n \|x_i - \tilde{\mathscr{S}}(z_i)\|_X^2 + \gamma \|\tilde{\mathscr{S}}\|_{\mathrm{HS}}^2 = \sum_{k=1}^\infty \left[\sum_{i=1}^n (\langle x_i, e_k \rangle_X - \langle z_i, \tilde{\mathscr{S}}^*(e_k) \rangle_Z)^2 + \gamma \|\tilde{\mathscr{S}} e_k\|^2\right].$$

Since each of the terms in square brackets can be chosen independently of each other, we have

$$\hat{\beta}_k := \hat{\mathscr{S}}_\gamma^*(e_k) = \operatorname*{argmin}_{\beta \in \mathcal{H}_Z} \sum_{i=1}^n (\langle x_i, e_k \rangle_X - \langle z_i, \beta \rangle_Z)^2 + \gamma \|\beta\|_Z^2.$$

A bit of matrix calculus combined with Lemma 2.16 yields that

$$(\langle z_1, \hat{\beta}_k \rangle_Z, \ldots, \langle z_n, \hat{\beta}_k \rangle_Z)^\top = K(K + \gamma I)^{-1} X_k^{(n)},$$

where I is the $n \times n$ identity matrix and $X_k^{(n)} := (\langle x_1, e_k \rangle_X, \ldots, \langle x_n, e_k \rangle_X)^\top$. Defining $\beta_k := \mathscr{S}^*(e_k)$, we can write $\beta_k = u_k + v_k$ where $u_k \in \mathcal{U} := \operatorname{span}(z_1, \ldots, z_n)$ and $v \in \mathcal{U}^\perp$. Writing $u_k = \sum_{j=1}^n \alpha_{k,j} z_j$ where $\alpha_k = (\alpha_{k,1}, \ldots, \alpha_{k,n})^\top \in \mathbb{R}^n$, we have for $i \in \{1, \ldots, n\}$,

$$\langle z_i, \beta_k \rangle_Z = \langle z_i, u_k \rangle_Z = \left\langle z_i, \sum_{j=1}^n \alpha_{k,j} z_j \right\rangle_Z = \sum_{j=1}^n \alpha_{k,j} \langle z_i, z_j \rangle_Z.$$

This entails

$$(\langle z_1, \beta_k\rangle_Z, \ldots, \langle z_n, \beta_k\rangle_Z)^\top = K\alpha_k.$$

Let $K = UDU^\top$ be the eigendecomposition of K, where $D_{ii} = \hat{\mu}_i$, and let $\theta_k := U^\top K\alpha_k$. Let $\varepsilon_k^{(n)} := (\langle \varepsilon_1, e_k\rangle_X, \ldots, \langle \varepsilon_n, e_k\rangle_X)^\top \in \mathbb{R}^n$ and note that $X_k^{(n)} = K\alpha_k + \varepsilon_k^{(n)}$. Letting $\|\cdot\|_2$ denote the Euclidean norm, n times the left-hand side of equation (2.40) can now be written (using equation (2.41))

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{\infty}\|K(K+\gamma I)^{-1}(U\theta_k + \varepsilon_k^{(n)}) - U\theta_k\|_2^2 \,|\, Z^{(n)}\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{\infty}\|DU^\top(UDU^\top + \gamma I)^{-1}(U\theta_k + \varepsilon_k^{(n)}) - \theta_k\|_2^2 \,|\, Z^{(n)}\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{\infty}\|D(D+\gamma I)^{-1}(\theta_k + U^\top\varepsilon_k^{(n)}) - \theta_k\|_2^2 \,|\, Z^{(n)}\right] \\
&= \sum_{k=1}^{\infty}\|(D(D+\gamma I)^{-1} - I)\theta_k\|_2^2 + \mathbb{E}\left[\sum_{k=1}^{\infty}\|D(D+\gamma I)^{-1}U^\top\varepsilon_k^{(n)}\|_2^2 \,|\, Z^{(n)}\right]
\end{aligned} \tag{2.42}$$

where the final equality uses that the first term is a function of $Z^{(n)}$ and the conditional expectation of the cross term in the sum of squares is 0, since $\mathbb{E}(\varepsilon_k^{(n)} \,|\, Z^{(n)}) = 0$.

The second term of (2.42) may be simplified as follows:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{\infty}\|D(D+\gamma I)^{-1}U^\top\varepsilon_k^{(n)}\|_2^2 \,|\, Z^{(n)}\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{\infty}\operatorname{tr}\left(D(D+\gamma I)^{-1}U^\top\varepsilon_k^{(n)}(\varepsilon_k^{(n)})^\top UD(D+\gamma I)^{-1}\right) \,|\, Z^{(n)}\right] \\
&= \operatorname{tr}\Bigg(D(D+\gamma I)^{-1}U^\top \underbrace{\mathbb{E}\left[\sum_{k=1}^{\infty}\varepsilon_k^{(n)}(\varepsilon_k^{(n)})^\top \,|\, Z^{(n)}\right]}_{\Sigma_{\varepsilon\,|\,Z}} UD(D+\gamma I)^{-1}\Bigg),
\end{aligned}$$

where we have used that only $\varepsilon_k^{(n)}$ is not a function of $Z^{(n)}$ and linearity of conditional expectations and the trace. Note that $\Sigma_{\varepsilon\,|\,Z}$ is a diagonal matrix with ith diagonal entry equal to

$$\mathbb{E}\left[\sum_{k=1}^{\infty}\langle \varepsilon_i, e_k\rangle_X^2 \,|\, Z^{(n)}\right] = \mathbb{E}\left[\|\varepsilon_i\|_X^2 \,|\, z_i\right],$$

hence we can bound each diagonal term by σ^2 by assumption. This implies that

$$\begin{aligned}
\operatorname{tr}\Big(D(D+\gamma I)^{-1}U^\top\Sigma_{\varepsilon\,|\,Z}UD(D+\gamma I)^{-1}\Big) &\le \sigma^2\operatorname{tr}\Big(D(D+\gamma I)^{-1}D(D+\gamma I)^{-1}\Big) \\
&= \sigma^2\sum_{i=1}^{n}\frac{\hat{\mu}_i^2}{(\hat{\mu}_i+\gamma)^2}.
\end{aligned}$$

The first term of (2.42) can be dealt with by noting that

$$\sum_{k=1}^{\infty}\|(D(D+\gamma I)^{-1}-I)\theta_k\|_2^2 = \sum_{k=1}^{\infty}\sum_{i=1}^{n}\frac{\gamma^2\theta_{k,i}^2}{(\hat{\mu}_i+\gamma)^2} = \sum_{k=1}^{\infty}\sum_{i:\hat{\mu}_i>0}\frac{\gamma^2\theta_{k,i}^2}{(\hat{\mu}_i+\gamma)^2}$$
$$= \sum_{k=1}^{\infty}\sum_{i:\hat{\mu}_i>0}\frac{\theta_{k,i}^2}{\hat{\mu}_i}\frac{\gamma^2\hat{\mu}_i}{(\hat{\mu}_i+\gamma)^2} \leq \left(\max_{i\in 1,\dots,n}\frac{\gamma^2\hat{\mu}_i}{(\hat{\mu}_i+\gamma)^2}\right)\sum_{k=1}^{\infty}\sum_{i:\hat{\mu}_i>0}\frac{\theta_{k,i}^2}{\hat{\mu}_i} \leq \frac{\gamma}{4}\sum_{k=1}^{\infty}\sum_{i:\hat{\mu}_i>0}\frac{\theta_{k,i}^2}{\hat{\mu}_i}.$$

The second equality uses that $\theta_k = U^\top K\alpha_k = DU^\top\alpha_k$, hence $\theta_{k,i}=0$ whenever $\hat{\mu}_i=0$ and the final inequality uses that $ab^2/(a+b)^2 \leq b/4$. Let D^+ denote the generalised inverse of D, i.e. $D^+_{ii} := \hat{\mu}_i^{-1}\mathbb{1}_{\hat{\mu}_i>0}$. Then

$$\sum_{i:\hat{\mu}_i>0}\frac{\theta_{k,i}^2}{\hat{\mu}_i} = \|\sqrt{D^+}\theta_k\|_2^2 = \alpha_k^\top KUD^+U^\top K\alpha_k = \alpha_k^\top UDD^+DU^\top\alpha_k = \alpha_k^\top K\alpha_k$$
$$= \|u_k\|_Z^2 \leq \|u_k\|_Z^2 + \|v_k\|_Z^2 = \|\beta_k\|_Z^2.$$

Putting things together, we have

$$\sum_{k=1}^{\infty}\|(D(D+\gamma I)^{-1}-I)\theta_k\|_2^2 \leq \frac{\gamma}{4}\sum_{k=1}^{\infty}\|\beta_k\|_Z^2 = \frac{\gamma}{4}\|\mathscr{S}\|_{\mathrm{HS}}^2.$$

Hence,

$$\frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}\|\mathscr{S}(Z_i)-\hat{\mathscr{S}}(Z_i)\|_Z^2 \mid Z^{(n)}\right) \leq \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\hat{\mu}_i^2}{(\hat{\mu}_i+\gamma)^2} + \frac{\gamma}{4n}\|\mathscr{S}\|_{\mathrm{HS}}^2,$$

and using that

$$\frac{\hat{\mu}_i^2}{(\hat{\mu}_i+\gamma)^2} \leq \min(1, \hat{\mu}_i^2/(4d_i\gamma)) = \min(\hat{\mu}_i/4, \gamma)/\gamma,$$

we have shown equation (2.40). □

To go from a conditional statement to an unconditional result, we first require the following lemma.

Lemma 2.18. *Let $x_1,\dots,x_n$ be i.i.d. observations of a centred Hilbertian random variable X with $E\|X\|^2<\infty$. Let $\mathscr{C}$ denote the covariance operator of X with eigen-expansion*

$$\mathscr{C} = \sum_{k=1}^{\infty}\mu_k e_k\otimes e_k \tag{2.43}$$

for an orthonormal basis $(e_k)_{k=1}^{\infty}$, and summable eigenvalues $\mu_1\geq\mu_2\geq\cdots\geq 0$. Define the random matrix $K\in\mathbb{R}^{n\times n}$ with entries given by $K_{ij}=\langle x_i,x_j\rangle$ and denote the eigenvalues of K/n by $\hat{\mu}_1\geq\hat{\mu}_2\geq\cdots\geq\hat{\mu}_n\geq 0$.

For all $r>0$,

$$\mathbb{E}\Big(\sum_{k=1}^{n}\min(\hat{\mu}_k, r)\Big) \leq \sum_{k=1}^{\infty}\min(\mu_k, r).$$

Proof. It suffices to show that given any $\epsilon > 0$, we have

$$\mathbb{E}\Big(\sum_{k=1}^{n} \min(\hat{\mu}_k, r)\Big) \leq \epsilon + \sum_{k=1}^{\infty} \min(\mu_k, r).$$

Now let d be such that

$$\sum_{k=d+1}^{\infty} \mu_k < \epsilon/n.$$

Let $\Phi \in \mathbb{R}^{n\times d}$ have entries given by

$$\Phi_{ij} := \langle x_i, e_j \rangle,$$

such that

$$(\Phi\Phi^\top)_{ij} := \sum_{k=1}^{d} \langle x_i, e_k\rangle\langle x_j, e_k\rangle.$$

From this, it is clear that

$$(K - \Phi\Phi^\top)_{ij} = \sum_{k=d+1}^{\infty} \langle x_i, e_k\rangle\langle x_j, e_k\rangle.$$

Thus, for $v \in \mathbb{R}^d$

$$v^\top(K - \Phi\Phi^\top)v = \sum_{i=1}^{d}\sum_{j=1}^{d} v_i v_j \sum_{k=d+1}^{\infty} \langle s_i, e_k\rangle\langle x_j, e_k\rangle = \sum_{k=d+1}^{\infty} \Big\langle \sum_{i=1}^{d} v_i x_i, e_k \Big\rangle^2 \geq 0,$$

showing that $K - \Phi\Phi^\top$ is positive semi-definite.

Next let $\mathbb{S}_+^d$ be the cone of positive semi-definite $d \times d$ matrices, and for $A \in \mathbb{S}_+^d$ and $k = 1, \ldots, d$, let $\lambda_k(A)$ denote the kth largest eigenvalue. Let $f : \mathbb{S}_+^d \to \mathbb{R}$ be given by

$$f(A) = \sum_{k=1}^{d} \min(\lambda_k(A), r).$$

By Weyl's inequality, noting that the non-zero eigenvalues of $\Phi^\top\Phi$ and $\Phi\Phi^\top$ coincide, we have for all k,

$$\hat{\mu}_k \leq \lambda_k(\Phi^\top\Phi/n) + \lambda_1(K - \Phi\Phi^\top/n)$$

and so

$$\min(\hat{\mu}_k, r) \leq \min(\lambda_k(\Phi^\top\Phi/n), r) + \mathrm{tr}(K - \Phi\Phi^\top)/n.$$

Thus,

$$\mathbb{E}\Big(\sum_{k=1}^{n} \min(\hat{\mu}_k, r)\Big) \leq \mathbb{E}f(\Phi^\top\Phi/n) + \mathbb{E}\mathrm{tr}(K - \Phi\Phi^\top). \tag{2.44}$$

Now by Fubini's theorem,

$$\mathbb{E}\mathrm{tr}(K - \Phi\Phi^\top) = \sum_{i=1}^{n} \sum_{k=d+1}^{\infty} \mathbb{E}(\langle x_i, e_k\rangle^2) = n \sum_{k=d+1}^{\infty} \mu_k < \epsilon.$$

We now claim that f is concave, from which the result will follow. Indeed, then by Jensen's inequality, $\mathbb{E}f(\Phi^\top\Phi/n) \leq f(\mathbb{E}\Phi^\top\Phi/n)$ and

$$\frac{1}{n}(\mathbb{E}\Phi^\top\Phi)_{kl} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\langle x_i, e_k\rangle\langle x_i, e_l\rangle) = \mu_k \mathbb{1}_{\{k=l\}}.$$

Thus,

$$f(\mathbb{E}\Phi^\top\Phi/n) = \sum_{k=1}^{d} \min(\mu_k, r),$$

and so returning to (2.44) we would have

$$\mathbb{E}\Big(\sum_{k=1}^{n} \min(\hat{\mu}_k, r)\Big) \leq \epsilon + \sum_{k=1}^{\infty} \min(\mu_k, r).$$

We now show that f is concave. Take $t \in (0,1)$ and $A, B \in \mathbb{S}_+^d$. We will show that

$$\sum_{k=1}^{d} (\lambda_k(tA + (1-t)B) - r)_+ \leq \sum_{k=1}^{d} \{t(\lambda_k(A) - r)_+ + (1-t)(\lambda_k(B) - r)_+\}, \tag{2.45}$$

where $(\cdot)_+$ denotes the positive part. This will prove concavity of f as

$$\begin{aligned}\sum_{k=1}^{d} \lambda_k(tA + (1-t)B) &= \mathrm{tr}(tA + (1-t)B) \\ &= t\mathrm{tr}(A) + (1-t)\mathrm{tr}(B) = \sum_{k=1}^{d} \{t\lambda_k(A) + (1-t)\lambda_k(B)\},\end{aligned}$$

so subtracting (2.45) yields $f(tA + (1-t)B) \geq tf(A) + (1-t)f(B)$ as desired.

Certainly (2.45) holds when $r \geq \lambda_1(tA + (1-t)B)$. Now by Lidskii's inequality, for each $j = 1, \ldots, d$,

$$\sum_{k=1}^{j} \lambda_k(tA + (1-t)B) \leq \sum_{k=1}^{j} \{t\lambda_k(A) + (1-t)\lambda_k(B)\}. \tag{2.46}$$

For convenience, let us set $\lambda_{d+1}(tA + (1-t)B) = 0$. Then for any $j = 1, \dots, d$, if $\lambda_{j+1}(tA + (1-t)B) \le r \le \lambda_j(tA + (1-t)B)$, we have

$$\begin{aligned}\sum_{k=1}^{d}(\lambda_k(tA + (1-t)B) - r)_+ &= \sum_{k=1}^{j}(\lambda_k(tA + (1-t)B) - r) \\ &\le \sum_{k=1}^{j}\{t(\lambda_k(A) - r) + (1-t)(\lambda_k(B) - r)\} \\ &\le \sum_{k=1}^{d}\{t(\lambda_k(A) - r)_+ + (1-t)(\lambda_k(B) - r)_+\},\end{aligned}$$

using (2.46) for the first inequality. We thus have that (2.45) holds whatever the value of r, and so f is concave, which completes the proof. □

Combining Lemma 2.17 and Lemma 2.18 now yields the following bound on our regression estimator.

Lemma 2.19. *Let $\mathcal{P}$ consist of a family of distributions of $(X, Z) \in \mathcal{H}_X \times \mathcal{H}_Z$ such that*

$$X = \mathscr{S}_P Z + \varepsilon_P,$$

where we assume that $\sup_{P\in\mathcal{P}}\|\mathscr{S}_P\|_{HS} < C$ and $\sup_{P\in\mathcal{P}}\mathbb{E}_P\|\varepsilon_P\|^2 < \sigma^2$. Suppose we are given n i.i.d. observations $(x_i, z_i)_{i=1}^n$ of (X, Z) and denote by $(\mu_{k,P})_{k\in\mathbb{N}}$ the non-negative eigenvalues of $\mathrm{Cov}_P(\varepsilon_P)$. Let $\mathscr{S}_\gamma$ be the estimator in (2.19). We have for each $P \in \mathcal{P}$, that

$$\frac{1}{n}\mathbb{E}_P\left(\sum_{i=1}^{n}\|\mathscr{S}(z_i) - \hat{\mathscr{S}}_\gamma(z_i)\|^2\right) \le \frac{\sigma^2}{\gamma}\frac{1}{n}\sum_{k=1}^{\infty}\min(\mu_{k,P}/4, \gamma) + \|\mathscr{S}_P\|_{HS}^2\frac{\gamma}{4n}. \tag{2.47}$$

Further, if we use $\hat{\gamma}$ as in (2.20), that is,

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma>0}\left(\frac{1}{\gamma n}\sum_{k=1}^{n}\min(\hat{\mu}_k/4, \gamma) + \frac{\gamma}{4}\right),$$

to produce an estimate $\hat{\mathscr{S}} := \hat{\mathscr{S}}_{\hat{\gamma}}$ of $\mathscr{S}_P$, then

$$\sup_{P\in\mathcal{P}}\mathbb{E}_P\left(\frac{1}{n}\sum_{i=1}^{n}\|\mathscr{S}_P(Z_i) - \hat{\mathscr{S}}(Z_i)\|^2\right) \le \max(\sigma^2, C)\sup_{P\in\mathcal{P}}\inf_{\gamma>0}\left(\frac{1}{\gamma n}\sum_{k=1}^{\infty}\min(\mu_{k,P}, \gamma) + \gamma\right). \tag{2.48}$$

Proof. Result (2.47) follows immediately from Lemmas 2.17 and 2.18. To show (2.48), we argue as follows. Let $(e_k)_{k\in\mathbb{N}}$ denote a basis of $\mathcal{H}_X$. Then conditioning on $z_1, \dots, z_n$ and applying

equation (2.40) in Lemma 2.17, we get that

$$\sup_{P\in\tilde{\mathcal{P}}_0} \mathbb{E}_P\left(\frac{1}{n}\sum_{i=1}^n \|\mathscr{S}_P(z_i)-\hat{\mathscr{S}}(z_i)\|^2\right) \leq \sup_{P\in\tilde{\mathcal{P}}_0} \mathbb{E}_P\left(\frac{\sigma^2}{\hat{\gamma}}\frac{1}{n}\sum_{k=1}^n \min(\hat{\mu}_k/4,\hat{\gamma}) + \|\mathscr{S}_P\|_{\mathrm{HS}}^2\frac{\hat{\gamma}}{4}\right)$$
$$\leq \max(\sigma^2, C)\sup_{P\in\tilde{\mathcal{P}}_0} \mathbb{E}_P\left[\min_{\gamma>0}\left(\frac{1}{\gamma n}\sum_{k=1}^n \min(\hat{\mu}_k/4,\gamma)+\frac{\gamma}{4}\right)\right].$$

Using the fact that the expectation of a minimum is less than the minimum of the expectation, we get that

$$\sup_{P\in\tilde{\mathcal{P}}_0} \mathbb{E}_P\left[\min_{\gamma>0}\left(\frac{1}{\gamma n}\sum_{k=1}^n \min(\hat{\mu}_k/4,\gamma)+\frac{\gamma}{4}\right)\right] \leq \sup_{P\in\tilde{\mathcal{P}}_0}\inf_{\gamma>0}\left[\mathbb{E}_P\left(\frac{1}{\gamma n}\sum_{k=1}^n \min(\hat{\mu}_k/4,\gamma)+\frac{\gamma}{4}\right)\right]$$
$$\leq \sup_{P\in\tilde{\mathcal{P}}_0}\inf_{\gamma>0}\left(\frac{1}{\gamma n}\sum_{k=1}^\infty \min(\mu_{k,P},\gamma)+\gamma\right),$$

where the second inequality is due to Lemma 2.18. □

Finally, we can prove Theorem 2.5.

Proof. By Theorem 2.3 and the assumptions of the Theorem it is sufficient to show that

$$\sup_{P\in\tilde{\mathcal{P}}_0} \sqrt{n}\mathbb{E}_P\left(\frac{1}{n}\sum_{i=1}^n \|\mathscr{S}_P^X(z_i)-\hat{\mathscr{S}}(z_i)\|^2\right) \to 0 \tag{2.49}$$

and similarly for the regression of Y on Z. This can be seen by noting that an application of Cauchy–Schwarz and Markov's inequality yields that $nM_{n,P}^f M_{n,P}^g \overset{P}{\rightrightarrows} 0$ and, by the upper bound on u_P and v_P in assumption (ii), $\tilde{M}_{n,P}^f \overset{P}{\rightrightarrows} 0$ and $\tilde{M}_{n,P}^g \overset{P}{\rightrightarrows} 0$.

Lemma 2.19 implies that it is sufficient to show that

$$\sqrt{n}\sup_{P\in\tilde{\mathcal{P}}_0}\inf_{\gamma>0}\left(\frac{1}{\gamma n}\sum_{k=1}^\infty \min(\mu_{k,P},\gamma)+\gamma\right) \to 0$$

as $n\to\infty$ for (2.49) to hold. For each $P\in\tilde{\mathcal{P}}_0$, we let $\phi_P:\mathbb{R}_+\to\mathbb{R}_+$ be given by

$$\phi_P(\gamma) = \sum_{k=1}^\infty \min(\mu_{k,P},\gamma).$$

By assumption (iii), $\lim_{\gamma \downarrow 0} \sup_{P \in \tilde{\mathcal{P}}_0} \phi_P(\gamma) = 0$, hence for any $\epsilon > 0$ we can find $N \in \mathbb{N}$ such that for any $n \geq N$, $\sup_{P \in \tilde{\mathcal{P}}_0} \sqrt{\phi_P\left(n^{-1/2}\right)} < \epsilon/2$. Let $\gamma_{n,P} = n^{-1/2}\sqrt{\phi_P\left(n^{-1/2}\right)}$. Then,

$$\sqrt{n} \sup_{P \in \tilde{\mathcal{P}}_0} \inf_{\gamma > 0} \left(\frac{1}{\gamma n} \sum_{k=1}^{\infty} \min(\mu_{k,P}, \gamma) + \gamma \right) = \sup_{P \in \tilde{\mathcal{P}}_0} \inf_{\gamma > 0} \left(\frac{\phi_P(\gamma)}{\gamma\sqrt{n}} + \sqrt{n}\gamma \right)$$

$$\leq \sup_{P \in \tilde{\mathcal{P}}_0} \left(\frac{\phi_P(\gamma_{n,P})}{\gamma_{n,P}\sqrt{n}} + \sqrt{n}\gamma_{n,P} \right) = \sup_{P \in \tilde{\mathcal{P}}_0} \left(\frac{\phi_P\left(n^{-1/2}\sqrt{\phi_P\left(n^{-1/2}\right)}\right)}{\sqrt{\phi_P\left(n^{-1/2}\right)}} + \sqrt{\phi_P\left(n^{-1/2}\right)} \right).$$

Assuming that $\epsilon \leq 2$ and using that ϕ_P is increasing, we get that for $n \geq N$,

$$\sup_{P \in \tilde{\mathcal{P}}_0} \left(\frac{\phi_P\left(n^{-1/2}\sqrt{\phi_P\left(n^{-1/2}\right)}\right)}{\sqrt{\phi_P\left(n^{-1/2}\right)}} + \sqrt{\phi_P\left(n^{-1/2}\right)} \right) < \sup_{P \in \tilde{\mathcal{P}}_0} \left(\frac{\phi_P\left(n^{-1/2}\epsilon/2\right)}{\sqrt{\phi_P\left(n^{-1/2}\right)}} + \sqrt{\phi_P\left(n^{-1/2}\right)} \right)$$

$$< \sup_{P \in \tilde{\mathcal{P}}_0} 2\sqrt{\phi_P\left(n^{-1/2}\right)} < \epsilon,$$

proving the result. □

Corollary 2.2. *Consider the setup of Lemma 2.19 but with the additional assumption that for some $a, b > 0$, we have $\mu_{k,P} \leq ae^{-bk}$ for all $P \in \mathcal{P}$. Then*

$$\sup_{P \in \tilde{\mathcal{P}}_0} \mathbb{E}_P \left(\frac{1}{n} \sum_{i=1}^{n} \|\mathscr{S}_P(z_i) - \hat{\mathscr{S}}(z_i)\|^2 \right) = o(\log n/n)$$

Proof. Applying Lemma 2.19, we show that

$$\sup_{P \in \tilde{\mathcal{P}}_0} \inf_{\gamma > 0} \left(\frac{1}{\gamma n} \sum_{k=1}^{\infty} \min(\mu_{k,P}, \gamma) + \gamma \right) \leq \inf_{\gamma > 0} \left(\frac{1}{\gamma n} \sum_{k=1}^{\infty} \min(ae^{-bk}, \gamma) + \gamma \right) = o(\log n/n).$$

To that end, note that

$$\frac{1}{\gamma n} \sum_{k=1}^{\infty} \min(ae^{-bk}, \gamma) + \gamma \leq -\frac{1}{nb} \log(\gamma/a) + \frac{1}{n\gamma} \int_{-\log(\gamma/a)/b}^{\infty} ae^{-xb}\, \mathrm{d}x + \gamma = -\frac{1}{nb} \log(\gamma/a) + \frac{1}{nb} + \gamma.$$

The right-hand side is a strictly convex function in γ hence it has a unique minimum at the unique root of the derivative function given by $\gamma^* := \frac{1}{nb}$ which yields a minimum of

$$\frac{1}{nb}(\log(anb) + 2) = o(\log n/n).$$

□

2.10 Additional numerical results

Here we include additional results relating to the setups in Section 2.5. Figures 2.7, 2.8 and 2.9 plot rejection rates against nominal significance levels for `pfr` and the GHCM, for the setups described in 2.5.1.

Figure 2.10 plots rejection rates for a subset of null settings considered in Section 2.5.1 but where the noise N_Y in (2.25) is t-distributed.

Figure 2.11 plots rejection rates for a subset of null settings considered in Section 2.5.1 but where instead of (2.25), the regression model for Y is given by

$$Y = \int_0^1 \alpha_a(t) Z(t) \mathrm{d}t + \sqrt{\frac{100}{n}} \int_0^1 \frac{\alpha_a(t)}{a} X(t) \mathrm{d}t + N_Y.$$

Note that when $n = 100$, the model is identical to (2.25). For other n, $\|\mathbb{E}\mathrm{Cov}(X, Y \mid Z)\|_{\mathrm{HS}}$ scales with $1/\sqrt{n}$, and so Theorem 2.4 suggests as n changes, the power should not change much. This is confirmed by our empirical results where we observe that the power remains largely unchanged as n changes, suggesting in particular that the GHCM has power against $1/\sqrt{n}$ alternatives.

Figure 2.12 plots rejection rates for the same settings considered in Section 2.5.1 but where we use the `FDboost` package for regressions instead of the `refund` package. We use default tuning parameters for the regression; it is possible that performance could improve with more careful tuning.

Figure 2.13 plots rejection rates for the same settings considered in Section 2.5.1 but where the X and Y curves are observed on an irregular grid with points sampled independently and uniformly on $[0, 1]$. We consider a sparse grid of 4 points as well as four unequal grid sizes sampled as the maximum of 4 and a Poisson random variable with mean in $\{10, 25, 50, 100\}$.

Figure 2.14 plots rejection rates for a simulation based on the real data analysis in Section 2.5.3. For each of the two edges that had Benjamini–Hochberg-corrected p-values at most 5% (O-L—PO-L and O-R—PO-R), we created artificial datasets as follows. We added independent Brownian motion noise to each of the estimated regression functions (note there were regression functions estimated for each variable in each of the two groups) thereby simulating a new X and Y conditional on the fixed Z. In these simulated datasets, the null of conditional independence does hold, and so we should expect the GHCM to deliver uniformly-distributed p-values. The results using the GHCM as described in Section 2.5.3 and for varying standard deviation σ of the Brownian motion noise for one set of regressions with the other set at 1, are shown in Figure 2.14. We see that even in the low σ settings, which are expected to be the most challenging, the GHCM maintains level control.

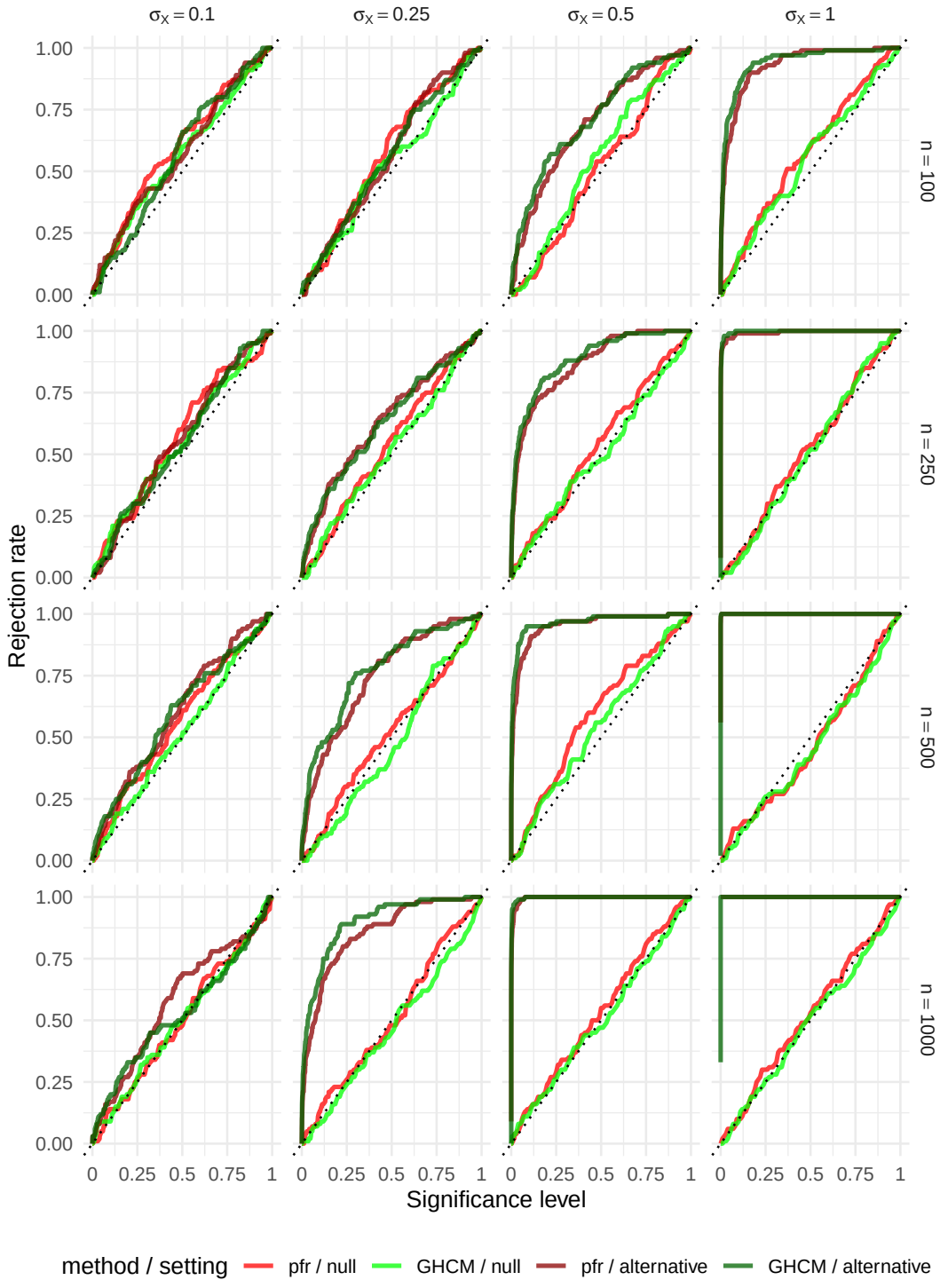


Fig. 2.7 Rejection rates against significance level α for the `pfr` (red) and GHCM (green) tests under null (light) and alternative (dark) settings when $a = 2$.

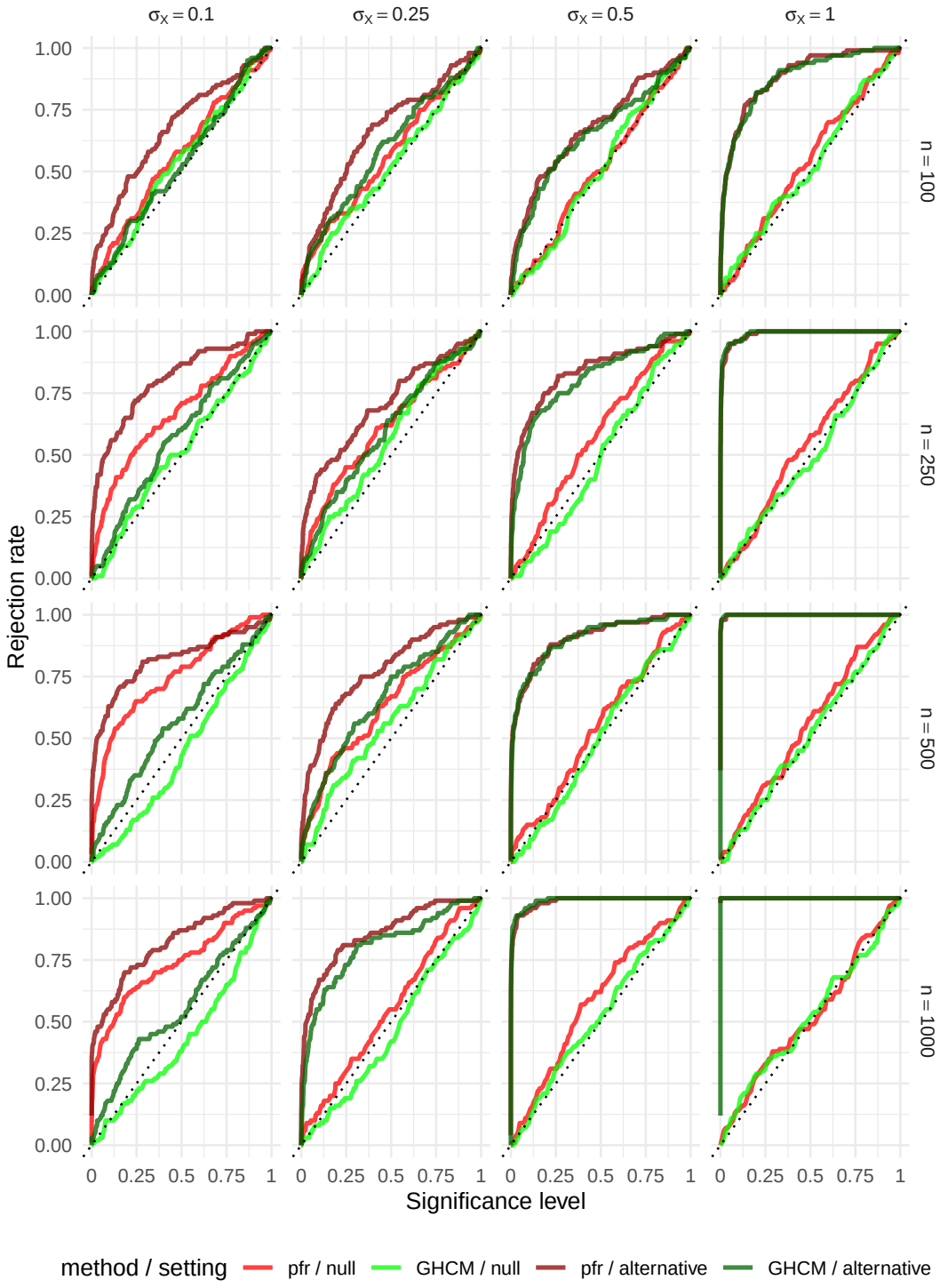


Fig. 2.8 Rejection rates against significance level α for the `pfr` (red) and GHCM (green) tests under null (light) and alternative (dark) settings when $a = 6$.

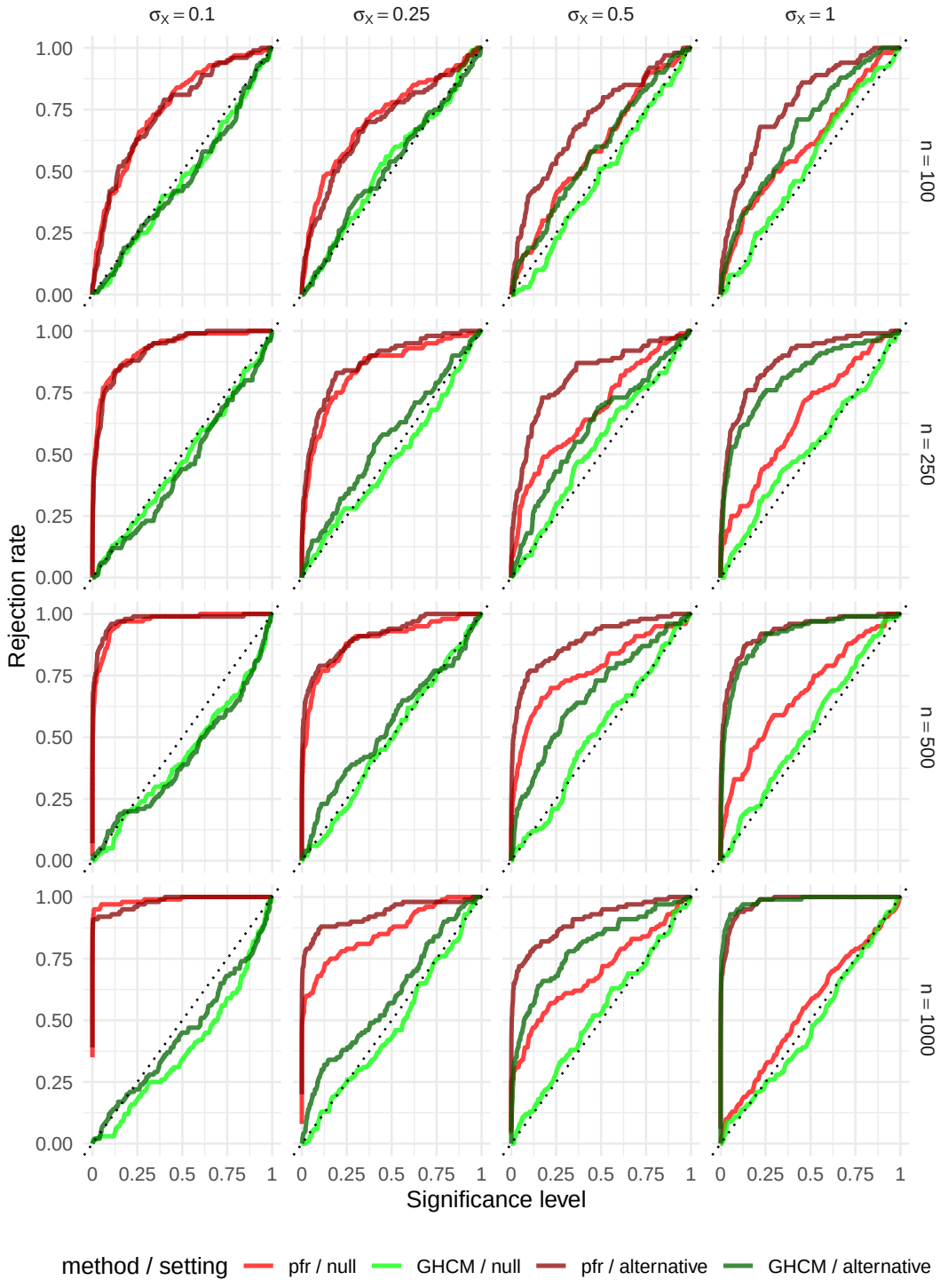


Fig. 2.9 Rejection rates against significance level α for the `pfr` (red) and GHCM (green) tests under null (light) and alternative (dark) settings when $a = 12$.

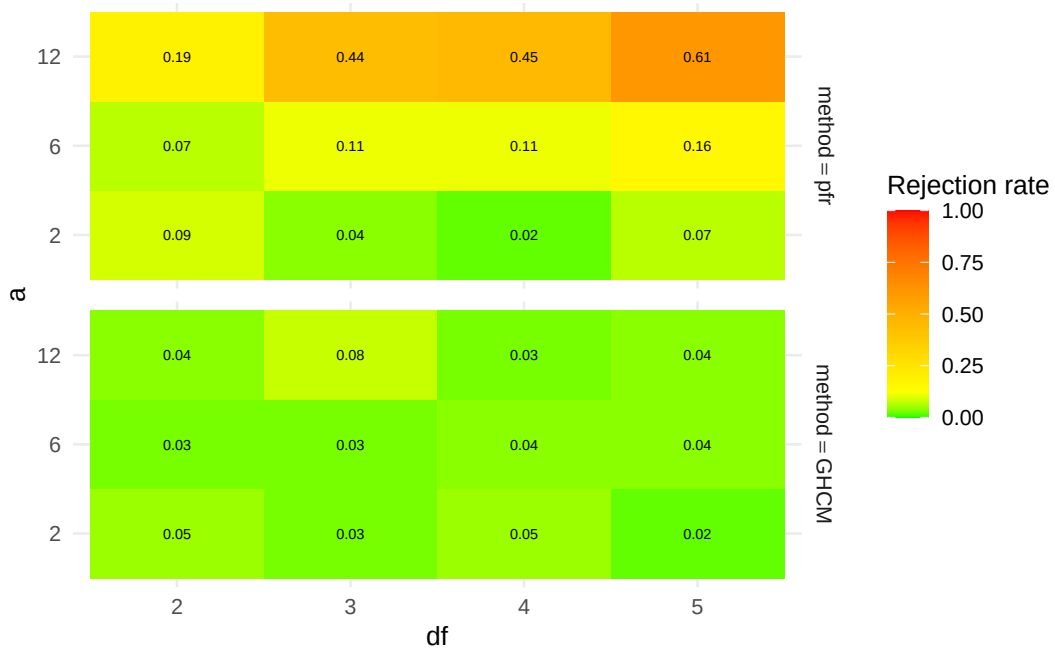


Fig. 2.10 Rejection rates in a subset of the null settings considered in Section 2.5.1 for the nominal 5%-level `pfr` test (top) and GHCM test (bottom) where $\sigma_X = 0.25$ and $n = 500$ and the noise N_Y in (2.25) is t-distributed with df degrees of freedom.

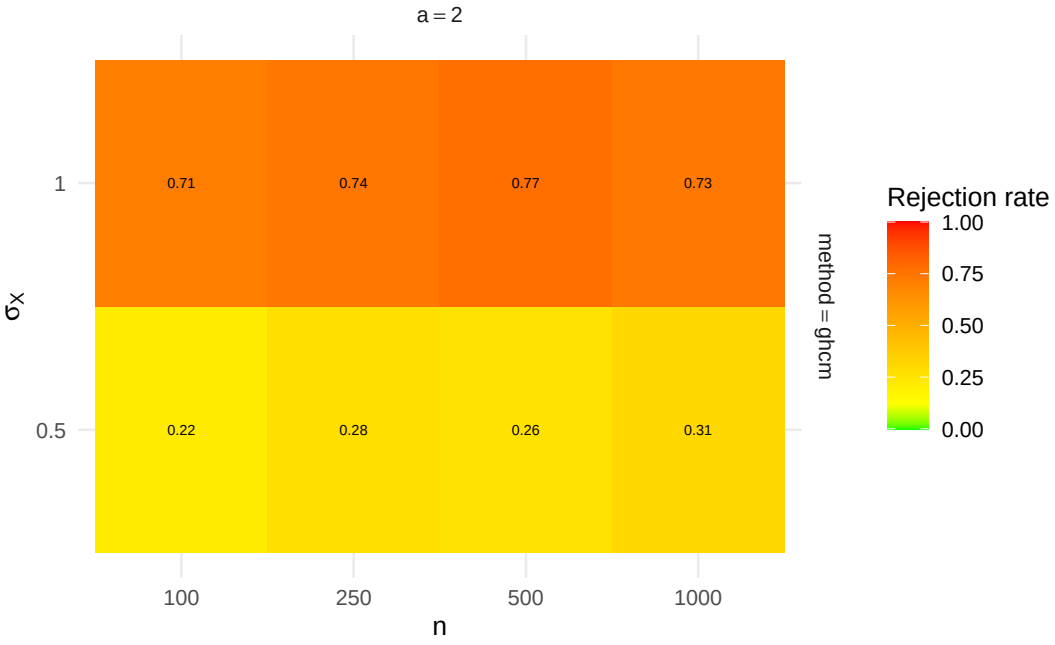


Fig. 2.11 Rejection rates in a subset of the alternative settings considered in Section 2.5.1 for the nominal 5%-level GHCM test where $a = 2$ and α_a has been replaced with $(100/n)^{-1/2}\alpha_a$.

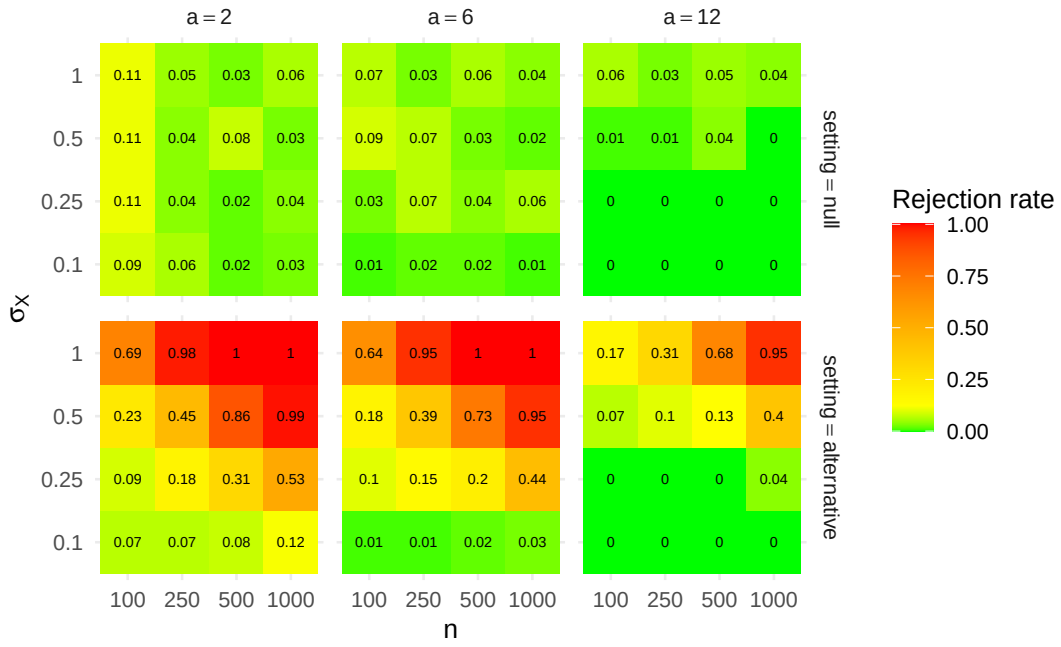


Fig. 2.12 Rejection rates in the setting of Section 2.5.1, replicating Figures 2.1 and 2.2, for the nominal 5%-level GHCM test using `FDboost` package for regressions instead of the `refund` package.

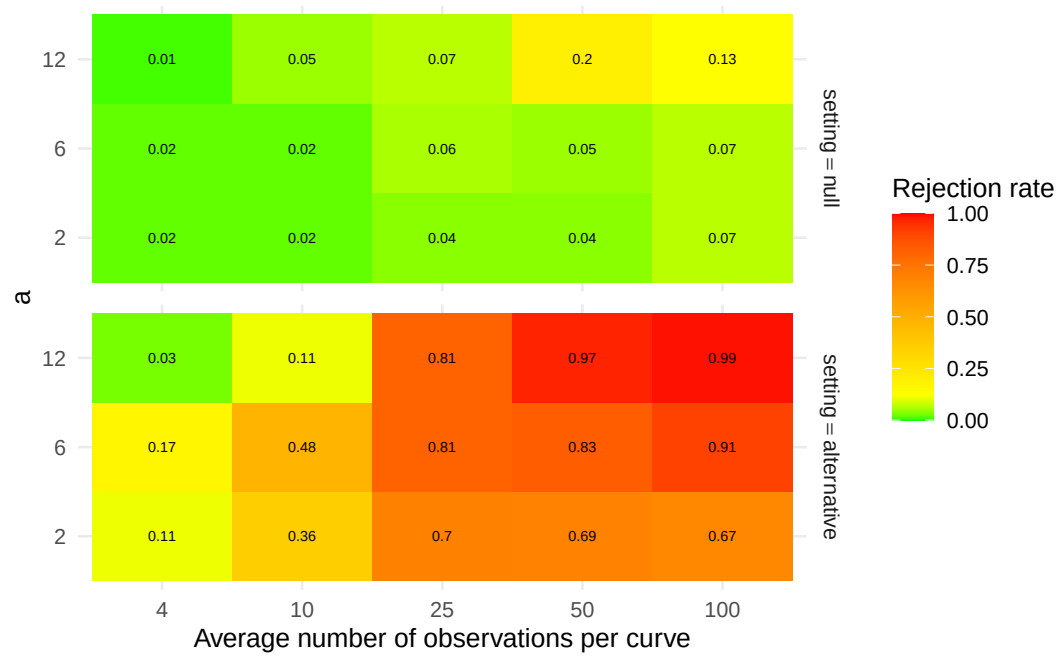


Fig. 2.13 Rejection rates in the setting of Section 2.5.1, replicating Figure 2.4, for the nominal 5%-level GHCM test where the X and Y curves are observed on irregular grids as described in the main text.

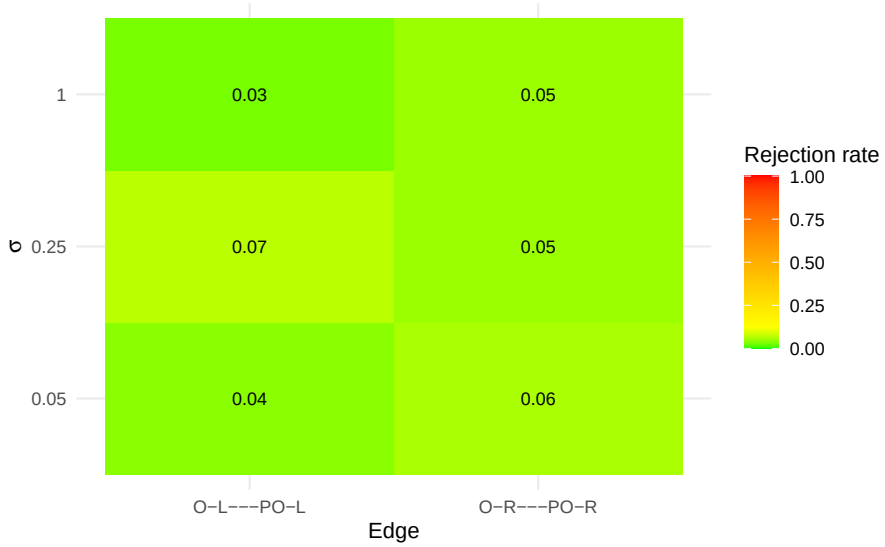


Fig. 2.14 Rejection rates for nominal 5%-level GHCM tests in simulation settings based on the EEG data studied in Section 2.5.3; see the main text for further details.

Chapter 3

The Projected Covariance Measure for model-free variable significance testing

3.1 Introduction

Understanding the relationship between a response and associated predictors is one of the most common problems faced by data analysts across many diverse areas of science and industry. Often an important step in this task is to determine which variables or groups of variables are important in this relationship. To fix ideas, consider data formed of independent copies of a triple (X, Y, Z), where $Y \in \mathbb{R}$ is our response, and we wish to assess the significance of a group of predictors $X \in \mathbb{R}^{d_X}$ after adjusting for confounding variables $Z \in \mathbb{R}^{d_Z}$; we will consider a more general setting later in this chapter where X and Z can be potentially non-Euclidean. One simple but popular way of addressing this problem is to fit a linear regression model $Y = X^T\beta + Z^T\gamma + \varepsilon$, where we assume the random error ε satisfies $\mathbb{E}(\varepsilon \,|\, X, Z) = 0$, and perform an F-test for the significance of X (i.e. test the null hypothesis that $\beta = 0$). However, in the case that the linear model is not a sufficiently good approximation of the ground truth, this can result in either wrongly declaring X to be important or unimportant, and other significance tests based on parametric models suffer from similar issues. The fact that regressions based on parametric models are greatly outperformed by modern machine learning methods such as deep learning (Goodfellow et al., 2016) and XGBoost (Chen and Guestrin, 2016) in regression competitions such as those hosted by Kaggle (Bojer and Meldgaard, 2021), suggests that such parametric models giving poor approximations to the truth is the norm rather than the exception, at least in contemporary datasets of interest.

In this work we consider the model-free null hypothesis of conditional mean independence, that is $\mathbb{E}(Y \,|\, X, Z) = \mathbb{E}(Y \,|\, Z)$; in words, X does not feature in the regression function of Y on X and Z. It is interesting to compare this to the conditional independence null $Y \perp\!\!\!\perp X \,|\, Z$, which has attracted much attention in recent years. The latter asks not just for the regression function to be expressed as a function of Z alone, but also the entire conditional distribution

of Y given (X, Z) to in fact equal the conditional distribution Y given Z. Any valid test of conditional mean independence may be used as a test for conditional independence as its size is no larger than its size over the larger null hypothesis of conditional mean independence. The two nulls in fact coincide when Y is binary, but more generally there are important differences. One attractive property of the conditional mean independence null is that the alternative of conditional mean dependence may be characterised by the property that X can improve the prediction of Y in a mean-squared error sense, given knowledge of Z. For example, consider the setting where X is a binary treatment variable, Z contains all pre-treatment confounders and Y is the observed outcome. Under assumptions (including the absence of unmeasured confounders) that are standard in the causal inference literature (Neyman, 1923; Rubin, 1974), conditional mean dependence is equivalent to the existence of a subgroup average treatment effect, that is a (measurable) subset $\mathcal{A} \subseteq \mathbb{R}^{d_Z}$ where $\mathbb{E}\{\mathbb{E}(Y \mid Z, X = 1) \mid Z \in \mathcal{A}\} > \mathbb{E}\{\mathbb{E}(Y \mid Z, X = 0) \mid Z \in \mathcal{A}\}$. On the other hand, rejection of the conditional independence null does not in general have an immediate interpretation in terms of its predictive implications.

Despite the attractions of conditional mean independence, an important issue is that this property is not testable without further restrictions on the null hypothesis: if (X, Y, Z) have a density that is absolutely continuous with respect to Lebesgue measure, then the power of any test at any alternative is at most its size. This comes as a direct consequence of the untestability of the smaller conditional independence null (Shah and Peters, 2020). The conclusion is that in order to test conditional mean independence, one must further constrain the null hypothesis in some way.

Given the success of machine learning methods in prediction problems, a natural and convenient way to specify these constraints is based on restricting the set of nulls to those where user-chosen regression methods can estimate certain conditional expectations sufficiently well. One strategy, as adopted in the *Generalised Covariance Measure* (GCM) of Shah and Peters (2020), involves, in the case where X is univariate, regressing each of X and Y on Z, computing the covariance between the resulting residuals and estimating a normalised version of $\mathbb{E}\{\mathrm{Cov}(X, Y \mid Z)\}$, a quantity that is zero under conditional independence. A drawback of this approach, however, is that it has no power against alternatives where $\mathbb{E}\{\mathrm{Cov}(X, Y \mid Z)\} = 0$.

To gain greater power, it is suggested to apply the above with X replaced by each component of $(\phi_1(X, Z), \ldots, \phi_m(X, Z))$, where $\phi_1, \ldots, \phi_m : \mathbb{R}^{d_X \times d_Z} \to \mathbb{R}$ are a fixed user-chosen collection of transformations of the data. One may then base a final test on the maximum absolute value of the resulting test statistics. It is however not clear how one should choose these transformations, and if m is large, or indeed d_X is large and we use the strategy above but with the ϕ_j simply extracting the jth component of X, performing all the regressions involved may be impractical. A related approach to improve the power properties of the GCM is introduced by Scheidegger et al. (2021) who propose a carefully weighted version of the GCM that, under conditions, can have power against alternatives where we do not have $\mathrm{Cov}(X, Y \mid Z) = 0$ almost surely, see also Fernández and Rivera (2022). Nevertheless, it is perfectly possible to have $\mathrm{Cov}(X, Y \mid Z) = 0$ under conditional mean dependence, and here even the weighted GCM would be powerless: for example, consider the simple setting where $(X, Z, \varepsilon) \sim N(0, \boldsymbol{I}_3)$ and $Y = X^2 + \varepsilon$. In this case, $\mathrm{Cov}(X, Y \mid Z) = \mathrm{Cov}(X, Y) = 0$ despite X clearly being important for the prediction of Y. It is

therefore of great interest to develop methods for testing conditional mean independence whose validity, as in the case of the GCM and its weighted version, relies primarily on the predictive properties of user-chosen regression methods, but have power against much wider classes of alternatives.

While there has been much work on the problem of conditional independence testing in recent years (we review some of the contribution most relevant to our work here in Section 3.1.2), there has been comparatively little on testing conditional mean independence. One compelling approach is based on an equivalent way of stating the null hypothesis: defining

$$\tau := \mathbb{E}\big[\{\mathbb{E}(Y \mid X, Z) - \mathbb{E}(Y \mid Z)\}^2\big] \tag{3.1}$$

we have that $\tau = 0$ if and only if we have conditional mean independence. This suggests a potential strategy for assessing conditional mean independence via the estimation of τ. Such an approach was adopted by Williamson et al. (2021) who employed a plug-in estimator of τ, and showed that, under conditions, it yields a semiparametric efficient estimator of τ, provided that $\tau > 0$. However, as highlighted by Williamson et al. (2021), under the null where $\tau = 0$, semiparametric approaches such as this face a fundamental difficulty as the influence function is identically zero, and as a consequence the test statistic has a degenerate distribution.

To avoid this issue, Williamson et al. (2022) and independently Dai et al. (2021), utilise an alternative representation of the target parameter $\tau = \mathbb{E}[\{Y - \mathbb{E}(Y \mid Z)\}^2] - \mathbb{E}[\{Y - \mathbb{E}(Y \mid X, Z)\}^2]$ and propose a testing procedure via sample splitting where estimation of $\mathbb{E}[\{Y - \mathbb{E}(Y \mid Z)\}^2]$ and $\mathbb{E}[\{Y - \mathbb{E}(Y \mid X, Z)\}^2]$ is done on independent splits of the data. This restores asymptotic normality of the test statistic under the null, but comes with a significant power loss. In particular, the resulting test becomes asymptotically powerless if $\sqrt{n}\tau \to 0$, even for a parametric linear model where the optimal testing rate is known to be of order n^{-1}. Moreover, the asymptotic normality fails when Y is (close to) independent of (X, Z), which raises concerns about uniform validity of the test. See Appendix 3.9.2 for details.

3.1.1 Outline of our approach and contributions

In view of the considerations above, the goal of this chapter is to propose a new framework for testing conditional mean independence that has the following properties.

- **Flexible type I error control.** The user should be able to leverage modern regression methods to ensure validity of the test uniformly over classes of distributions where these methods perform sufficiently well.
- **Adaptive power.** The test should have minimax rate-optimal power in both simple parametric models and challenging nonparametric settings, when used with appropriate regression methods.
- **Practical.** The test should involve only performing a small number of regressions, so it is computationally practical.

Our approach is based on the following alternative characterisation of conditional mean independence: Y is conditionally mean independent of X given Z if and only if

$$\mathbb{E}[\{Y - \mathbb{E}(Y \mid Z)\} f(X, Z)] = \mathbb{E}[\mathrm{Cov}(Y, f(X, Z) \mid Z)] = 0 \tag{3.2}$$

for all functions f such that $\mathbb{E}(f(X, Z)^2) < \infty$. In words, the residuals $Y - \mathbb{E}(Y \mid Z)$ from regressing Y on Z alone are uncorrelated with any function of X and Z. On the other hand, under an alternative, these residuals should not be pure noise but contain some 'signal' that can be exposed via an appropriate f such that the left-hand side of (3.2) is strictly positive.

To motivate our approach, consider an oracular test statistic that uses knowledge of the conditional expectation $\mathbb{E}(Y \mid Z)$: given independent copies $(X_i, Y_i, Z_i)_{i=1}^n$ of (X, Y, Z) and a function f, the random variables $L_i^* := \{Y_i - \mathbb{E}(Y_i \mid Z_i)\} f(X_i, Z_i)$ for $i = 1, \ldots, n$ are independent and identically distributed, with zero mean under the null. Writing $\widetilde{L}_i^* := \{Y_i - \mathbb{E}(Y_i \mid X_i, Z_i)\} f(X_i, Z_i)$, we have that under regularity conditions, the studentised statistic

$$T^* := \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n L_i^*}{\sqrt{\frac{1}{n} \sum_{i=1}^n \widetilde{L}_i^{*2}}} \tag{3.3}$$

converges to a standard normal distribution under the null, and may thus form the basis of a test. Note that since under the null, $\widetilde{L}_i^* = L_i^*$, we may alternatively studentise the test statistic using the empirical standard deviation of the L_i^*; however this version simplifies the derivation to follow.

Different choices of f would lead to different power properties under an alternative. Ideally, we want to maximise the value of the test statistic under an alternative, so we would like $\mathbb{E}(L_i^*)/\sqrt{\mathrm{Var}(\widetilde{L}_i^*)}$ to be as large as possible. It may be shown (see Proposition 3.5 in Section 3.9.2 of the appendix) that this is uniquely maximised, up to an arbitrary positive scaling, by choosing $f(X, Z) = h(X, Z)/v(X, Z)$, where $h(X, Z) := \mathbb{E}(Y \mid X, Z) - \mathbb{E}(Y \mid Z)$ and $v(X, Z) := \mathrm{Var}(Y \mid X, Z)$. We therefore see that the optimal f is a version of the projection h of Y onto the space of square-integrable functions of (X, Z) that are orthogonal to functions of Z, inversely weighted by the conditional variance v.

The considerations above suggest the following approach: use one portion of the data to obtain an estimate of the projection f, and then use the remaining data to evaluate a test statistic of the form (3.3). This forms the basis of our proposed test statistic, which we call the *Projected Covariance Measure* (PCM).

One important issue to be addressed is the fact that under the null, h is the zero function, and as a consequence, both the numerator and denominator of T^* are zero. This is not immediately problematic for the oracular statistic T^*, as one can always decide to accept the null when the numerator is precisely 0. However, it might appear to be potentially disastrous for an empirical version of T^* where any bias terms in the numerator could be inflated by division with a denominator that is close to zero. One of our main contributions in this work is to show that by formulating our PCM test statistic appropriately, it has an asymptotic standard Gaussian limit in settings ranging from low- and high-dimensional linear models to

fully nonparametric settings. Moreover, we demonstrate empirically that this limiting behaviour can be expected to hold more generally when using machine learning methods such as random forests (Breiman, 2001) for the regressions involved.

The rest of the chapter is organised as follows. After reviewing some related literature in Section 3.1.2, we present our PCM methodology in Section 3.2. In Section 3.3, we examine the simplest instantiation of our general framework and study testing in the context of low-dimensional linear models. An important revelation of this analysis is that in contrast to the equally general testing frameworks of Williamson et al. (2022) and Dai et al. (2021), our approach has power against local alternatives where τ is of order n^{-1}. We go on to show that, under conditions, the PCM maintains Type I error control in high-dimensional linear models, even when using an essentially arbitrary machine learning method to estimate the projection f. In Section 3.4, we present a general theory of the PCM, giving conditions involving prediction errors of the user-chosen regression procedures used in the PCM that result in Type I error control, and also study the power of the procedure. In Section 3.5, we show how our general conditions for Type I error control may be satisfied in a fully nonparametric regression setting when using series estimators for the relevant regressions. We also introduce a slight variant of our approach involving additional sample splitting that enjoys what we show to be minimax rate optimal power over classes of alternatives for which τ in (3.1) satisfies a lower bound.

All of our asymptotic results are uniform, in the sense that they give classes of distributions over which the probabilities of rejecting the null hypothesis are simultaneously controlled. As discussed earlier, no non-trivial test of conditional mean independence can maintain its nominal level over the entirety of the null. Understanding the classes of null distributions over which Type I error may be controlled for a given test therefore becomes crucially important.

In Section 3.6, we conduct several simulation experiments that demonstrate the effectiveness of the PCM when used with generalised additive model-based regressions (Wood, 2017) and random forests, in terms of both Type I error control and power. We conclude with a discussion in Section 3.7 outlining potential future research directions suggested by our work.

In Section 3.8 and 3.9 of the appendix, we include the proofs of all our main results and related auxiliary lemmas. Section 3.10 provides a self-contained description of spline regression and related results that we use for our analysis in Section 3.5. In Section 3.11, we give a more detailed analysis of our results for linear projections in Section 3.3; in particular we derive an exact asymptotic power function of our test. Section 3.12 contains the results from additional numerical experiments beyond those included in Section 3.6.

3.1.2 Literature review

There is a relatively small body of literature that is explicitly concerned with conditional mean independence. Early developments on this topic include the work of Fan and Li (1996), Lavergne and Vuong (2000) and Aït-Sahalia et al. (2001) from the econometrics community. Jin et al. (2018) propose an approach for testing conditional mean independence in cases where $\mathbb{E}(Y \mid Z)$ is a linear function of Z, based on the martingale difference divergence proposed by Shao and Zhang (2014).

Recent years have witnessed an increasing use of machine learning (ML) tools for statistical inference. For example, Chernozhukov et al. (2018) introduce an ML-driven approach for estimating causal parameters in the presence of complex nuisance parameters. Shah and Bühlmann (2018) and Janková et al. (2020) propose methods for goodness-of-fit testing in high-dimensional (generalised) linear models that involve detecting remaining signal in residuals using ML methods. More closely related to this work, Williamson et al. (2022), and independently Dai et al. (2021), propose model-free methods for assessing conditional mean independence that can take advantage of existing ML algorithms. Williamson et al. (2022) derive a semiparametrically efficient estimation of τ, but recognise the difficulty of testing the null hypothesis that $\tau = 0$ caused by the fact that the efficient influence function is identically zero under the null. This means that their sample-splitting approach lacks validity when (X, Y, Z) are independent, and moreover it turns out the test may require larger values of τ than necessary in order to achieve power; see Section 3.3.1 for a more detailed discussion. Dai et al. (2021) alleviate the Type I error issue by adding noise to their test statistic, but this comes at a further price in terms of power, as pointed out by Verdinelli and Wasserman (2021). We also mention the work of Zhang and Janson (2020), who provide a method of constructing confidence intervals for τ in the case where the conditional distribution of X given Z is (almost) known.

Many existing tests, including ours, determine their critical values based on asymptotic theory derived under the null. However, most work (implicitly) targets pointwise Type I error control that holds only each fixed null. This type of pointwise analysis leaves a room for the existence of a sequence of null distributions for which the Type I error can be made arbitrarily large. A classical example is the fact that the t-test that has pointwise asymptotic size α for the class of distributions with finite variance, but uniform asymptotic size 1 for the same class of distributions (Romano, 2004). In fact, even more seriously, if we let $\mathcal{P}$ denote the class of distributions on $\mathbb{R}$ with finite mean μ, and consider a random sample from some $P \in \mathcal{P}$, then no test of the null that $\mu = 0$ can have power greater than its size at *any* alternative (Bahadur and Savage, 1956; Romano, 2004). We therefore put great emphasis on uniform Type I error control over classes of distributions in order to present more practically-relevant error guarantees. This uniform analysis is in line with recent work on conditional independence testing such as Candès et al. (2018), Berrett et al. (2020), Shah and Peters (2020), Petersen and Hansen (2021), Lundborg et al. (2021a), Scheidegger et al. (2021) and Neykov et al. (2021).

Our work builds on a classical technique, namely sample splitting, that involves partitioning the data into disjoint subsamples for different purposes: roughly speaking, a portion of the data is used for seeking a good direction that potentially contains a high signal and the other portion is used for conducting a test based on the data projected along the given direction. Cox (1975) is one of the earliest papers that applies sample splitting to testing problems. Since then, many inference procedures have been developed by leveraging a similar technique to perform variable selection in high-dimensional models (Meinshausen and Bühlmann, 2010; Meinshausen et al., 2009; Shah and Samworth, 2013; Wasserman and Roeder, 2009), inference after model selection (Rinaldo et al., 2019) and inference based on maximum likelihood estimators (Wasserman et al., 2020), to name just a few. In a similar vein, Kim and Ramdas (2020) introduce splitting-based procedures that address an issue of degenerate U-statistics

for high-dimensional inference. While our main focus is on testing, sample splitting has also been considered for estimation problems, where it typically works as a device to reduce a bias and thus help to obtain a fast (often optimal) convergence rate (Chernozhukov et al., 2018; Newey and Robins, 2018; Wang and Shah, 2020). Some parts of our work are motivated by Newey and Robins (2018), who propose cross-fit estimators of functionals involving conditional expectations.

3.1.3 Preliminaries and notation

Throughout this chapter, we adopt the convention that $0/0 := 0$ and $\mathrm{sgn}(0) := 0$. We let $x \wedge y := \min(x, y)$ throughout and denote by $[n] := \{1, \ldots, n\}$. For two sequences (a_n) and (b_n), we write $a_n \asymp b_n$ to mean that there exist $c, C > 0$ such that $0 < c \leq |a_n/b_n| \leq C < \infty$ for every n. For a vector $x \in \mathbb{R}^n$ and $p \in [1, \infty]$, we denote its ℓ_p norm by $\|x\|_p$. The operator norm of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ is denoted by $\|\boldsymbol{A}\|_{\mathrm{op}}$. For a positive semi-definite matrix $\boldsymbol{A}$, we write $\boldsymbol{A}^{-1}$ for its generalised inverse. We use the notation z_α to denote the α quantile of the standard normal distribution whose cumulative distribution function is denoted by Φ.

In order to present our uniform results on testing, we require some conventions for probabilistic notation used in what follows. Let $(\Omega, \mathcal{F})$ be a measurable space equipped with a family of probability measures $(\mathbb{P}_P)_{P \in \mathcal{P}}$ where $\mathcal{P}$ is a collection of distributions on a Euclidean space. We will permit the family $\mathcal{P}$ to depend on n, to allow for settings where the number of parameters grows with n, but will typically suppress this in the notation.

Given a family of sequences of random variables $(X_{P,n})_{P \in \mathcal{P}, n \in \mathbb{N}}$ on $(\Omega, \mathcal{F})$ whose distributions are determined by $P \in \mathcal{P}$, we write $X_{P,n} = o_{\mathcal{P}}(1)$ if $\sup_{P \in \mathcal{P}} \mathbb{P}_P(|X_{P,n}| > \epsilon) \to 0$ for every $\epsilon > 0$. Similarly, we write $X_{P,n} = O_{\mathcal{P}}(1)$ if, for any $\epsilon > 0$, there exist $M_\epsilon, N_\epsilon > 0$ such that $\sup_{n \geq N_\epsilon} \sup_{P \in \mathcal{P}} \mathbb{P}_P(|X_{P,n}| > M_\epsilon) < \epsilon$. In addition, for another family of sequences of random variables $(Y_{P,n})_{P \in \mathcal{P}, n \in \mathbb{N}}$, we write $X_{P,n} = o_{\mathcal{P}}(Y_{P,n})$ and $X_{P,n} = O_{\mathcal{P}}(Y_{P,n})$ if there exists $R_{P,n}$ with $X_{P,n} = Y_{P,n} R_{P,n}$ and $R_{P,n} = o_{\mathcal{P}}(1)$ and $R_{P,n} = O_{\mathcal{P}}(1)$, respectively. We say that $(X_{P,n})_{P \in \mathcal{P}, n \in \mathbb{N}}$ converges uniformly in distribution to random variable X with distribution function F if for all x where F is continuous,

$$\lim_{n \to \infty} \sup_{P \in \mathcal{P}} \left| \mathbb{P}_P(X_{P,n} \leq x) - F(x) \right| = 0.$$

We will throughout denote different independent datasets by $\mathcal{D}_i$ for $i \in \mathbb{N}$, each containing n independent observations. In what follows, we often abuse notation and write conditional expectations conditioning on a random function, e.g. $\mathbb{E}_P(\widehat{f}(X, Z) \,|\, \widehat{f})$ where $\widehat{f}$ is a function produced by some regression estimator. By this we formally mean that we condition on the sample used to construct the regression estimator and any additional randomness involved in the computation of the regression function. Throughout the rest of the chapter we let (X, Y, Z) be random variables on $\mathcal{X} \times \mathbb{R} \times \mathcal{Z}$, although we will at times think of $\mathcal{X}$ and $\mathcal{Z}$ being specific d_X and d_Z-dimensional Euclidean spaces, respectively.

3.2 Projected covariance measure

In this section, we formally present our PCM methodology. There are some modifications we make to the basic blueprint outlined in Section 3.1.1, and we first motivate these in Section 3.2.1 before presenting our final algorithm in Section 3.2.2. Given that our approach involves sample splitting, it is convenient to assume here and also throughout Sections 3.3 and 3.4 that we have $2n$ i.i.d. observations $(X_i, Y_i, Z_i)_{i=1}^{2n}$ rather than the conventional n observations.

3.2.1 Motivation

Recall that the approach sketched in Section 3.1.1 involves first computing an estimate $\widehat{f}$ of the weighted projection

$$f(X,Z) = \frac{h(X,Z)}{v(X,Z)} = \frac{\mathbb{E}(Y \mid X, Z) - \mathbb{E}(Y \mid Z)}{\mathrm{Var}(Y \mid X, Z)}$$

using one portion of the data, say $\mathcal{D}_2 := (X_i, Y_i, Z_i)_{i=n+1}^{2n}$. We discuss how to construct the estimate $\widehat{f}$ in Section 3.2.2. Next, given an estimate $\widehat{m}(\cdot)$ of $m(\cdot) := \mathbb{E}(Y \mid Z = \cdot)$, the oracular test statistic (3.3) suggests a numerator of our test statistic of the form

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(Z_i)\} \widehat{f}(X_i, Z_i). \tag{3.4}$$

We would like this to have mean approximately zero under the null; however it is well-known (Chernozhukov et al., 2018) that when using a nonparametric estimator $\widehat{m}$, the quantity above may carry a substantial bias, and we should instead consider an orthogonalised version of the form

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_i \qquad \text{with} \qquad L_i := \{Y_i - \widehat{m}(Z_i)\}\{\widehat{f}(X_i, Z_i) - \widehat{m}_{\widehat{f}}(X_i, Z_i)\},$$

where $\widehat{m}_{\widehat{f}}$ is an estimate of $m_{\widehat{f}}(\cdot) := \mathbb{E}(\widehat{f}(X,Z) \mid Z = \cdot, \widehat{f})$. Importantly, the bias term then involves a product of the mean squared prediction error (MSPE) of $\widehat{m}$,

$$\frac{1}{n} \sum_{i=1}^{n} \{m(Z_i) - \widehat{m}(Z_i)\}^2, \tag{3.5}$$

and that of $\widehat{f}$, a quantity that may be substantially smaller than the MSPE of $\widehat{m}$ alone (which would drive the bias in (3.4)).

Turning to the denominator of our test statistic, rather than studentising by a quantity requiring an estimate of $\mathbb{E}(Y \mid X, Z)$ as suggested by (3.3), it is practically more convenient normalise using the empirical standard deviation of $L_1, \ldots, L_n$ as this does not involve performing an additional regression. Thus, we propose to take as our test statistic

$$T := \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} L_i}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} L_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} L_i^2\right)}}. \tag{3.6}$$

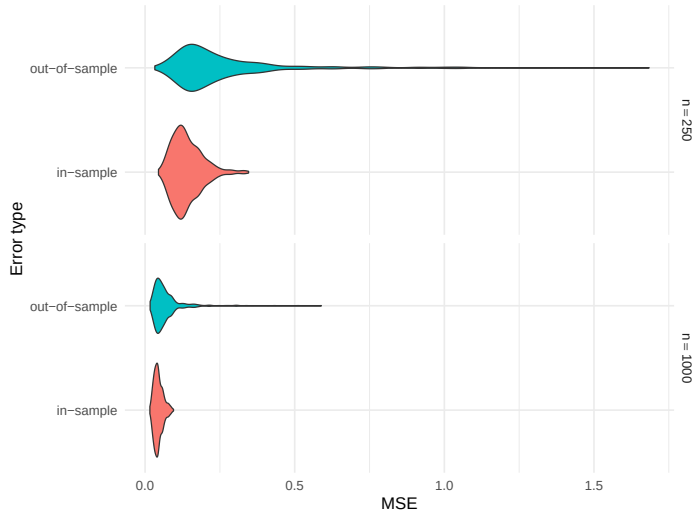


Fig. 3.1 In-sample and out-of-sample errors for the model where $Z_1, \dots, Z_7 \sim N_7(0, \boldsymbol{I})$, $Y = \sin(2\pi Z_1) + \varepsilon$ with $\varepsilon \sim N(0,1)$ independently of $Z_1, \dots, Z_7$, and regressions are performed using `mgcv`; see Section 3.6.1 for more details on this setup.

For local alternatives, both versions are near-identical and so any differences in power properties should be very slight, as we have also observed empirically. Recalling that the GCM is of the form (3.6) with $(L_i)_{i=1}^n$ equal to the product of the regression errors from regression each of Y and X onto Z, we see that our final test statistic is the GCM (Shah and Peters, 2020) applied to a transformed $X \mapsto \widehat{f}(X, Z)$, with the transformation chosen using $\mathcal{D}_2$ to maximise the power of the test.

As in the GCM, we choose in practice train $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ on $\mathcal{D}_1$ rather than $\mathcal{D}_2$. The errors such as (3.5) that are required to be controlled are then in-sample errors, that is the regression methods are trained on the same data they are evaluated on, and thus the regression methods need not extrapolate to unseen data points, for example. While from a theoretical perspective in-sample errors and out-of-sample errors are often thought of similarly, in finite samples, these can behave differently: for example Figure 3.1 demonstrates that when using additive models (computed using the `R` package `mgcv` (Wood, 2017)) to estimate m in a setup considered in Section 3.6.1, out-of-sample errors can be appreciably larger with non-negligible probability.

As the PCM may be thought of as the GCM applied to a transformed X, we would hope to obtain a standard Gaussian limit for T as in the case of the regular GCM test statistic. Given that the transformation is designed to result in large values of T under an alternative, we would perform a one-sided test by rejecting when T exceeds the appropriate normal quantile. Unfortunately however, the theory that guarantees asymptotic validity of the GCM test statistic does not apply in our case: it would require $\mathrm{Var}(\{Y - m(Z)\}\{\widehat{f}(X,Z) - m_{\widehat{f}}(Z)\} \mid \widehat{f})$, i.e. the (square of the) target of the denominator to be bounded away from zero under the null. However, f is identically 0 under the null so $\widehat{f}$ and hence the above variance, and also both numerator and denominator of our test statistic, should all converge to 0.

To see why we can expect a standard Gaussian limit for our test statistic despite this apparent degeneracy, consider a linear model setting where $(X, Y, Z) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ are related through

$$Y = \beta X + Z^\top \boldsymbol{\gamma} + \varepsilon \qquad \text{and} \qquad X = Z^\top \boldsymbol{\eta} + \xi, \tag{3.7}$$

with $\beta = 0$ and $\mathbb{E}(\varepsilon \,|\, Z) = \mathbb{E}(\xi \,|\, Z) = 0$. If we form estimates $\widehat{h}$ and $\widehat{m}$ using ordinary least squares, and for simplicity set $\widehat{v} \equiv 1$ when forming $\widehat{f}$, then $\widehat{f}(x, z) = \widehat{h}(x, z)$ takes the form $\widehat{\beta} x + z^\top \widetilde{\boldsymbol{\delta}}$ for some $(\widehat{\beta}, \widetilde{\boldsymbol{\delta}}) \in \mathbb{R} \times \mathbb{R}^d$. Note that both $\widehat{\beta}$ and $\|\widetilde{\boldsymbol{\delta}}\|_2$ are of stochastic order $1/\sqrt{n}$.

Let us write $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\eta}}$ for the regression estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ respectively. The next step of our procedure involves regressing each of $(Y_i)_{i=1}^n$ and $\big(\widehat{f}(X_i, Z_i)\big)_{i=1}^n$ onto $(Z_i)_{i=1}^n$. The residuals from the latter regression take the form $\widehat{\beta}\{Z_i^\top(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}) + \xi_i\}$, so in our case

$$L_i = \widehat{\beta}\{Z_i^\top(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}) + \varepsilon_i\}\{Z_i^\top(\boldsymbol{\eta} - \widehat{\boldsymbol{\eta}}) + \xi_i\}.$$

Thus, although L_i and hence its standard deviation would be of order $1/\sqrt{n}$ due to the factor of $\widehat{\beta}$, writing $L_i' := L_i/|\widehat{\beta}|$, we see that our test statistic is of the form $\mathrm{sgn}(\widehat{\beta})T'$, where T' is a version of T in (3.6) with L_i replaced by L_i'. But L_i' is an order 1 quantity (in contrast of L_i), so under mild conditions $n^{-1/2}\sum_{i=1}^n L_i'$ will have a non-degenerate Gaussian limit, yielding a standard Gaussian limit for T'. As $\widehat{\beta}$ is independent of T', having been constructed on $\mathcal{D}_2$, the final test statistic T will also converge to a standard Gaussian.

While this argument provides a heuristic justification for the asymptotic validity of our proposed test under a simple linear model, there remain challenges in extending the basic intuition of this example to more general settings. In the above, it was possible to isolate the randomness from $\widehat{f}$ simply via the sign of $\widehat{\beta}$, which helps bypass the $0/0$ issue. However, it is by no means straightforward to deal with the limits of the form $0/0$ in a nonparametric setting where $\widehat{f}$ is entangled with other sources of randomness in a complicated way. Moreover, in nonparametric settings one needs to put more effort into ensuring that the convergence rates of $\widehat{f}$, $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ are fast enough that the bias term is asymptotically negligible. In this process, we are obliged to handle a nested regression problem that has rarely been touched in the literature with a few exceptions (e.g. Kennedy, 2020).

3.2.2 PCM algorithm

Our PCM approach developed in Section 3.2.1 is set out in Algorithm 2, with some recommendations for the constructions of $\widehat{h}$ and $\widehat{v}$ which we discuss below in Sections 3.2.2 and 3.2.2. In Section 3.2.2, we then put forward a version of the PCM using multiple sample splits that we recommend using in practice.

Choice of $\widehat{h}$

We would like $\widehat{h}(X, Z)$ to be close to $h(X, Z) = \mathbb{E}(Y \,|\, X, Z) - \mathbb{E}(Y \,|\, Z)$ in order to maximise the power of the procedure. There are several ways of estimating h, perhaps the most obvious being simply to take the difference of the estimated regression functions $\widehat{g}$ and $\check{m}$ from regressing Y on each of (X, Z) and Z. An alternative approach is based on observing that

Algorithm 2 Projected Covariance Measure: single sample split

Input: Data $(X_i, Y_i, Z_i)_{i=1}^{2n}$, significance level $\alpha \in (0,1)$, partition of $[2n] = \mathcal{I}_1 \cup \mathcal{I}_2$ into index sets $\mathcal{I}_1$ and $\mathcal{I}_2$, each of size n.
Options: Regression methods for each of the regressions.
Define: $\mathcal{D}_j = (X_i, Y_i, Z_i)_{i \in \mathcal{I}_j}$ for $j \in [2]$.

1: *Form* $\widehat{h}$.

(i) Regress Y onto (X, Z) using $\mathcal{D}_2$ to give fitted regression function $\widehat{g}$.

(ii) If $\widehat{g}$ can be modified so that all components involving only Z are set to 0, let $\widetilde{g}$ be this modified version of $\widehat{g}$. Alternatively, set $\widetilde{g} := \widehat{g}$.

(iii) Regress $\widetilde{g}(X, Z)$ onto Z using $\mathcal{D}_2$ to give $\widetilde{m}$, and then set $\widetilde{h}(x, z) := \widetilde{g}(x, z) - \widetilde{m}(z)$.

(iv) Compute

$$\hat{\rho} := \frac{1}{n} \sum_{i \in \mathcal{I}_2} \{Y_i - \widehat{g}(X_i, Z_i) + \widetilde{g}(X_i, Z_i) - \widetilde{m}(Z_i)\} \widehat{h}(X_i, Z_i),$$

and set $\widehat{h}(x, z) := \operatorname{sgn}(\hat{\rho}) \widetilde{h}(x, z)$,

2: *Form* $\widehat{v}$.

(i) Regress $(Y - \widehat{g}(X, Z))^2$ onto (X, Z) using $\mathcal{D}_2$ to give $\widetilde{v}$.

(ii) Define $a : [0, \infty) \to [0, \infty]$ by

$$a(c) = \frac{1}{n} \sum_{i \in \mathcal{I}_2} \frac{\{Y_i - \widehat{g}(X_i, Z_i)\}^2}{\max(\widetilde{v}(X_i, Z_i), 0) + c}.$$

If $a(0) \leq 1$, set $\hat{c} := 0$; otherwise find $\hat{c}$ by solving $a(c) = 1$. Set $\widehat{v}(x, z) := \max(\widetilde{v}(x, z), 0) + \hat{c}$.

3: *Compute test statistic.*

(i) Set $\widehat{f}(x, z) := \widehat{h}(x, z)/\widehat{v}(x, z)$ and regress $\widehat{f}(X, Z)$ onto Z using $\mathcal{D}_1$, giving $\widehat{m}_{\widehat{f}}$.

(ii) Regress Y onto Z using $\mathcal{D}_1$ to give $\widehat{m}$.

(iii) For $i \in \mathcal{I}_1$ set $L_i := \{Y_i - \widehat{m}(Z_i)\}\{\widehat{f}(X_i, Z_i) - \widehat{m}_{\widehat{f}}(Z_i)\}$ and let

$$T := \frac{\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}_1} L_i}{\sqrt{\frac{1}{n} \sum_{i \in \mathcal{I}_1} L_i^2 - (\frac{1}{n} \sum_{i \in \mathcal{I}_1} L_i^2)}}.$$

4: *Reject* H_0 *if* $T > z_{1-\alpha}$ *otherwise accept* H_0.

$h(X,Z) = g(X,Z) - \mathbb{E}(g(X,Z) \mid Z)$ where $g(X,Z) := \mathbb{E}(Y \mid X,Z)$. This suggests subtracting not $\check{m}$ but the output of regressing $\widehat{g}(X,Z)$ onto Z. An advantage of this latter approach is that we are free to subtract any function r of Z from $\widehat{g}(X,Z)$ prior to this second regression onto Z, as we also have $h(X,Z) = g(X,Z) - r(Z) - \mathbb{E}(g(X,Z) - r(Z) \mid Z)$. Thus, for example if $\widehat{g}(x,z) = \widehat{g}_x(x) + \widehat{g}_z(z)$, we may form an estimate of $h(X,Z)$ as the residuals from regressing $\widehat{g}_x(X)$ onto Z. This second regression can then focus on removing any Z signal in $\widehat{g}_x(X)$, rather than also having to also having to cancel out $\widehat{g}_z(Z)$. We do not make the claim that this always makes a large improvement on the first approach, and indeed for certain regression methods such as ordinary least squares (OLS), both approaches are identical and the 'cancellation' is automatic. Nevertheless, we find the approach of Step 1 of Algorithm 2 a sensible default choice.

In Step 1 (iv) we make a final modification to the estimate thus constructed by potentially flipping its sign. The rationale for this is as follows: under an alternative, we have that $\mathbb{E}[\{Y - \mathbb{E}(Y \mid Z)\}h(X,Z)] = \tau > 0$. As a basic check then, we can see if an empirical version of this inequality, with $\widehat{h}$ taking place of h and an estimate of $\mathbb{E}(Y \mid Z)$ replacing the population quantity, holds; if not, we can at least flip the sign of $\widehat{h}$. This does not require performing any further regressions to estimate $\mathbb{E}(Y \mid Z)$: noting the identity $\mathbb{E}(Y \mid Z) = \mathbb{E}\{\mathbb{E}(Y \mid X,Z) - r(Z) \mid Z\} + r(Z)$, observe that $\widetilde{m}$ is an estimate of the first of these quantities taking $r(Z) = \widehat{g}(X,Z) - \widetilde{g}(X,Z)$; see 1 (ii) and (iii). We note that when using OLS for each of the regressions, $\hat{\rho}$ is guaranteed to be non-negative, so no sign flip is performed.

In high-dimensional settings, we would typically use a sparsity-inducing regression method such as the Lasso (Tibshirani, 1996). Considering the simple case where X is univariate, this can result in the coefficient for X being set exactly to zero, and so the recommended construction of $\widehat{h}$ given above would simply produce the zero function. Whilst not a problem for Type I error control, as our convention (see Section 3.1.3) is to accept when $L_i = 0$ for all i, it is wasteful in terms of power and a better approach here would be to leave the coefficient for X unpenalised. More generally for multivariate X, we can additionally regress on the first principal component of X for example, and leave this unpenalised.

Choice of $\widehat{v}$

A natural way of forming $\widehat{v}$ is via regressing the square of the residuals from regressing Y onto (X,Z) onto (X,Z), and this is what we recommend in Step 2 (i) of Algorithm 2 to produce $\widetilde{v}$. An issue is that whilst v is clearly non-negative, and expected to be positive everywhere, $\widetilde{v}$ may in fact be negative. In fact equally problematic is the possibility that $\widetilde{v}$ is very close to 0 at some (X_i,Z_i), as then taking $\widehat{v} = \widetilde{v}$, we would have $\widehat{f}(X_i,Z_i)$ very large and hence $\widehat{f}(X_i,Z_i) - \widehat{m}_{\widehat{f}}(Z_i)$ and L_i may be greatly inflated and dominate the test statistic. To mitigate these problems, we modify $\widetilde{v}$ by taking the positive part of our initial estimate, and then adding a non-negative constant $\hat{c}$. This constant is chosen such that $a(\hat{c})$ (see Step 2(ii) of Algorithm 2) is at most 1, the rationale for this coming from the population level identity $\mathbb{E}[\{Y - \mathbb{E}(Y \mid X,Z)\}^2/v(X,Z)] = 1$. We also note that estimation of the conditional variance v is not critical for good power properties. For example, in Section 3.5 we show that simply

setting $\widehat{v} \equiv 1$ delivers minimax rate optimal power in a fully nonparametric setting; however the power properties may improve empirically by a constant factor, see Section 3.3.1.

Multiple sample splitting

The single sample split in Algorithm 2 crucially ensures independence between $\widehat{f}$ and the remaining data $\mathcal{D}_1$, but has the consequence of introducing unwanted additional randomness to the test statistic. To mitigate this issue, we advocate applying the single split PCM to multiple splits of the data, and averaging the resulting test statistics, as summarised in Algorithm 3. An alternative to working with the averaged test statistic would be to combine the p-values of the individual tests, for which several methods are available ranging from twice the average or median of p-values to Bonferroni method (e.g. DiCiccio et al., 2020; Meinshausen et al., 2009; Vovk and Wang, 2020, and references therein). However, our experience is that these approaches tend to be overly conservative, and typically lose power compared to considering a single test. Instead, we propose to compare the averaged test statistic $\bar{T}$ to a standard Gaussian quantile, as with the single split test statistic T; a similar approach is taken in Wang and Shah (2020). We expect this to be conservative, as by Jensen's inequality, $\bar{T}$ is less than or equal to T in the convex ordering, so for example $\mathrm{Var}(\bar{T}) \leq \mathrm{Var}(T)$. However, in practice it does tend to improve slightly on the power of a single-split test, whilst importantly also derandomising it. It is worth noting that one could also do cross-fitting here, i.e. always apply the test to both $(\mathcal{I}_1^{(b)}, \mathcal{I}_2^{(b)})$ and $(\mathcal{I}_2^{(b)}, \mathcal{I}_1^{(b)})$ in Algorithm 3. We experienced no benefits in terms of either type I error control or power from doing this though.

Algorithm 3 Projected Covariance Measure: multiple sample splits

Input: Data $(X_i, Y_i, Z_i)_{i=1}^{2n}$, significance level $\alpha \in (0, 1)$, number of splits B.
Options: Regression methods for each of the regressions.

Test Statistic:

Form complementary pairs of index sets $\{(\mathcal{I}_1^{(b)}, \mathcal{I}_2^{(b)}) : b = 1, \ldots, B\}$ each of size n, where $\mathcal{I}_1^{(b)} \cup \mathcal{I}_2^{(b)} = [2n]$.

For each $b = 1, \ldots, B$, apply Algorithm 2 with index sets $\mathcal{I}_1^{(b)}, \mathcal{I}_2^{(b)}$ to produce test statistic $T^{(b)}$.

Return $\bar{T} := \sum_{b=1}^{B} T^{(b)}/B$.

Decision: **if** $\bar{T} > z_{1-\alpha}$ **then** reject H_0 **else** accept H_0.

As previously mentioned, this modified procedure seeks to derandomise the test to mitigate one of the primary downside of employing sample-splitting like we do: the resulting test is random so that two analyses of the same data can result in contradictory inference simply due to the choice of sample split. We discuss the issue of calibration above but there is also the obvious additional computational cost involved in repeated sample-splitting that can be substantial if B is large – our proposed procedure involves many regressions that could be time-consuming. We believe, however, that these downsides are outweighed by the advantages of increased power and de-randomisation.s

3.3 Linear models

In this section we study our PCM methodology in the context of a linear model for Y on X and Z. We begin with the simplest version of this setup, where we assume that $g(x,z) := \mathbb{E}(Y \mid X = x, Z = z)$ is a linear function that we estimate using ordinary least squares. This is not the sort of challenging setting where we would envision applying the PCM in practice, as clearly a t-test (modified to account for potential heteroscedasticity) would suffice to test for the significance of X. We nevertheless present it to show that in contrast to the general methodologies put forward by Williamson et al. (2022) and Dai et al. (2021), here our method has power against $n^{-1/2}$-alternatives. In Section 3.3.2 below, we show that for both low- and high-dimensional Z, we retain Type I error control even under an arbitrary model for X and when using an essentially arbitrary estimated projection $\widehat{f}$.

3.3.1 Linear projection function

Here we consider a family of $\mathcal{P}$ of joint distributions P of $(X, Y, Z) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ satisfying a linear model,

$$Y = \beta_P X + \boldsymbol{\gamma}_P^\top Z + \zeta, \tag{3.8}$$

where $\beta_P \in \mathbb{R}$ and $\boldsymbol{\gamma}_P \in \mathbb{R}^d$ are regression coefficients and ζ is a random noise term with $\mathbb{E}_P(\zeta \mid X, Z) = 0$. We further impose the following moment conditions on $\mathcal{P}$.

Assumption 3.1. Assume that the family $\mathcal{P}$ of joint distributions P of $(X, Y, Z) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ satisfies (3.8). Let $\boldsymbol{\eta}_P$ denote the population least squares projection of X onto Z and define $\xi_P := X - \boldsymbol{\eta}_P^\top Z$. Let $\boldsymbol{\Theta}_P := \mathbb{E}_P(ZZ^\top \xi_P^2)$, $W = (X, Z) \in \mathbb{R}^{d+1}$ and $\boldsymbol{\Sigma}_P^{XZ} := \mathbb{E}_P(WW^\top)$. Assume that there exist $C, c, \delta > 0$ such that

(i) $\sup_{P \in \mathcal{P}} \max\{\mathbb{E}_P(\|Z\|_2^{4+\delta}), \mathbb{E}_P(|Y|^{4+\delta}), \mathbb{E}_P(|X|^{4+\delta})\} \leq C$.

(ii) $\inf_{P \in \mathcal{P}} \min\{\mathrm{Var}_P(\zeta), \lambda_{\min}(\boldsymbol{\Sigma}_P^{XZ}), \lambda_{\min}(\boldsymbol{\Theta}_P)\} \geq c$.

Proposition 3.1. *Consider a version of the PCM setting $\widehat{v} \equiv 1$ and using OLS for each of the regressions involved for a family of distributions $\mathcal{P}$ satisfying Assumption 3.1. Let $\mathcal{P}_1(\kappa) := \{P \in \mathcal{P} : |\beta_P| = \kappa/\sqrt{n}\}$. Given any $\eta, \alpha \in (0,1)$, there exists $\kappa > 0$ such that*

$$\lim_{n \to \infty} \inf_{P \in \mathcal{P}_1(\kappa)} \mathbb{P}(T > z_{1-\alpha}) > \eta.$$

Proposition 3.1 gives the reassuring conclusion that in the simplest of settings, our general PCM framework, when used with appropriately chosen regression methods, can up to a constant match the power properties of a t-test tailored to this setting. The setting is in fact simple enough for us to derive an asymptotic power expression for our test. In Section 3.11 of the appendix we present such an analysis for a version of our test that uses n_1 and n_2 (with $n_1 + n_2 = 2n$) observations in $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, rather than the equal split we consider here. This shows that the optimal splitting ratio depends on the unknown signal strength, and

therefore supports a default choice of $n_1 = n_2 = n$ for simplicity. We also provide a simulation study in Section 3.12.1 of the appendix where we compare the local power properties of the PCM, the test byWilliamson et al. (2022) and the F-test with a robust variance estimator.

3.3.2 A general estimated projection

We next consider a situation where the model is unspecified under the alternative, whereas Y has a linear relationship with Z under the null of conditional mean independence. In this case, it is reasonable to employ a flexible regression method, such as neural networks or random forests, to estimate the projection f. Our goal here is to identify conditions on estimators, including $\widehat{f}$, under which the proposed test controls the Type I error. It turns out that, given a specified null model, the problem of testing whether X is significant is closely connected to goodness-of-fit testing for the null model, and we are able to exploit this connection to study the asymptotic Type I error of the proposed test.

Consider first the case of low-dimensional Z. Let $\mathcal{P}_0$ be a family of distributions of (X, Y, Z) under the null where $Z \in \mathbb{R}^d$ has an arbitrary distribution, and we suppose that $Y = \boldsymbol{\gamma}_P^\top Z + \varepsilon$, where $\mathbb{E}_P(\varepsilon \,|\, X, Z) = 0$. Therefore, $m_P(z) = \boldsymbol{\gamma}_P^\top Z$ and thus it is reasonable to use a linear regression model for $\widehat{m}$. We will suppose that $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ in Algorithm 2 are performed using OLS, whereas we will leave the regression choices involved in the construction of $\widehat{f}$ arbitrary. We define T_{OLS} to denote the resulting test statistic and make the following assumptions to ensure uniform asymptotic normality of the test statistic.

Assumption 3.2. Consider a class of null distributions $\mathcal{P}_0$ of (X, Y, Z) where $Y = \boldsymbol{\gamma}^\top Z + \varepsilon$ with $\mathbb{E}_P(\varepsilon \,|\, X, Z) = 0$ and assume that

(a) There exist $\delta \in (0, 2]$, $c, C > 0$ such that $\mathbb{E}_P(\varepsilon^2 \,|\, X, Z) \geq c$ and $\mathbb{E}_P(|\varepsilon|^{2+\delta} \,|\, X, Z) \leq C$ for all $P \in \mathcal{P}_0$.

(b) For $i \in [n]$, let $u_{n,i} := \widehat{f}(X_i, Z_i) - \widehat{m}_{\widehat{f}}(Z_i)$ and $v_{n,i} := u_{n,i} / \big(\sum_{i'=1}^n u_{n,i'}^2\big)^{1/2}$. We assume that $\max_{i\in[n]} |v_{n,i}| = o_{\mathcal{P}_0}(1)$ and $\sum_{i=1}^n v_{n,i}^2 = 1 + o_{\mathcal{P}_0}(1)$.

(c) Letting $\widehat{\boldsymbol{\gamma}}$ denote the coefficient from the $\widehat{m}$ regression, we assume that $\max_{i\in[n]} \|Z_i\|_\infty \cdot \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 = o_{\mathcal{P}_0}(1)$.

Part (a) of Assumption 3.2 concerns conditional moments of ε, and is used to establish the asymptotic normality of a suitably normalised In contrast to prior work on goodness-of-fit testing, e.g. Janková et al. (2020), we do not assume that the conditional variance of ε is homogeneous. Assumption 3.2(b) asks for no particular $|v_{n,i}|$ to be significantly larger than the others, and, for large enough n, that at least one of $\{u_{n,i} : i \in [n]\}$ is non-zero for all $P \in \mathcal{P}_0$. The latter condition is important for establishing the asymptotic normality of our test statistic, but not crucial for Type I error control. Indeed, when $u_{n,i} = 0$ for all $i \in [n]$, the test statistic is zero, and we do not reject the null. Finally, in settings where, for example each Z_i is sub-Gaussian with parameter 1, then $\max_{i\in[n]} \|Z_i\|_\infty \leq \sqrt{3\log(nd_Z)}$ and $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \lesssim n^{-1/2}$ with high probability, and in that case Part (c) of Assumption 3.2 is satisfied.

Proposition 3.2 (Low-dimensional Z)**.** *Suppose in the above setting that Assumption 3.2 holds and that all OLS estimators exist almost surely. Then the test statistic T_{OLS} converges to $N(0,1)$ uniformly over $\mathcal{P}_0$; i.e.,*

$$\sup_{P\in\mathcal{P}_0}\sup_{t\in\mathbb{R}}\left|\mathbb{P}_P(T_{\mathrm{OLS}}\leq t)-\Phi(t)\right|\to 0.$$

From Proposition 3.2, the test that rejects the null when $T_{\mathrm{OLS}} > z_{1-\alpha}$ is uniformly asymptotically level α under Assumption 3.2. We also note that the only requirement we impose on the projection $\widehat{f}$ is that it satisfies Assumption 3.2(b). Janková et al. (2020) also consider this condition, providing empirical supporting evidence in general, and introducing a specific procedure that guarantees that the condition holds.

We now extend these ideas and the setting described above Assumption 3.2 to the case where the dimension of Z is potentially larger than the sample size. Here, the least squares estimator is not necessarily well-defined, so to address this issue, we construct $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ using the Lasso or one of its variants. Letting $\widehat{\gamma}$ denote the coefficients from the $\widehat{m}$ regression, the motivation for this comes from the decomposition

$$\sum_{i=1}^{n}(Y_i-\widehat{\gamma}^\top Z_i)u_{n,i}=\sum_{i=1}^{n}\varepsilon_i u_{n,i}-\delta_{\mathrm{bias}}, \tag{3.9}$$

where $\delta_{\mathrm{bias}} := \sum_{i=1}^{n}(\widehat{\gamma}-\gamma)^\top Z_i u_{n,i}$. While this bias term is no longer exactly zero as for the least squares estimators considered in Proposition 3.2, Hölder's inequality will nevertheless guarantee that it is sufficiently small for our purposes as long as

$$\|\widehat{\gamma}-\gamma\|_1\cdot\left\|\sum_{i=1}^{n}Z_i v_{n,i}\right\|_\infty=o_{\mathcal{P}_0}(1). \tag{3.10}$$

We denote the test statistic as described in Algorithm 2 in this context by T_{Lasso}. The next proposition is the analogue of Proposition 3.2 for T_{Lasso}.

Proposition 3.3 (High-dimensional Z)**.** *Suppose in the above setting that Assumption 3.2 and condition (3.10) hold. Then*

$$\sup_{P\in\mathcal{P}_0}\sup_{t\in\mathbb{R}}\left|\mathbb{P}_P(T_{\mathrm{Lasso}}\leq t)-\Phi(t)\right|\to 0.$$

In order to ensure condition (3.10), one can use the square-root Lasso (Belloni et al., 2011), as suggested by Janková et al. (2020). In particular, for $\lambda_{\mathrm{sq}} > 0$, we set $\widehat{m}_{\widehat{f}}(z)=\widehat{\boldsymbol{\eta}}_{\mathrm{sq}}^\top z$ where

$$\widehat{\boldsymbol{\eta}}_{\mathrm{sq}}:=\operatorname*{argmin}_{\boldsymbol{\eta}\in\mathbb{R}^d}\left\{\frac{1}{\sqrt{n}}\|\widehat{\boldsymbol{f}}-\boldsymbol{Z}\boldsymbol{\eta}\|_2+\lambda_{\mathrm{sq}}\|\boldsymbol{\eta}\|_1\right\}.$$

With this choice of $\widehat{\boldsymbol{\eta}}_{\mathrm{sq}}$ and by letting $\lambda_{\mathrm{sq}}=C\sqrt{(\log d_Z)/n}$, the Karush–Kuhn–Tucker conditions for the square-root Lasso guarantee that $\|\sum_{i=1}^{n}Z_i v_{n,i}\|_\infty \leq C\sqrt{\log d_Z}$. Furthermore, under appropriate conditions, the Lasso estimator $\widehat{\gamma}$ has an error bound $\|\widehat{\gamma}-\gamma\|_1 \lesssim s_0\sqrt{\log d_Z/n}$

with high probability, where s_0 is the number of non-zero coefficients of $\boldsymbol{\gamma}$ (e.g. Corollary 6.2 of Bühlmann and van de Geer, 2011). Therefore, in this setting, condition (3.10) is satisfied if $s_0(\log d_Z)/\sqrt{n} \to 0$.

3.4 General theory

In this section, we present general conditions ensuring uniform asymptotic validity and power of the test, primarily by imposing assumptions on the performance of the regressions involved in computing our test. To facilitate our analysis, it is helpful to study a slight modification of the test as presented in Algorithm 2, where we form $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ on an independent auxiliary sample. In principle, we may accomplish this by further splitting $\mathcal{D}_1$ into two, and using one part to train $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$, and the other to compute the test statistic. Moreover, we can exchange the roles of the two parts and average the resulting test statistics, a process known as cross-fitting, which then guarantees no loss in efficiency from this additional sample split. However, for the reasons discussed in Section 3.2.1 we do not recommend performing this in practice.

The following quantities relating to the performances of the regression methods $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ will play a key role in our results. Let us introduce

$$\varepsilon_{P,i} := Y_i - m_P(Z_i), \qquad \xi_{P,i} := \widehat{f}(X_i, Z_i) - m_{P,\widehat{f}}(Z_i), \tag{3.11}$$

$$\sigma_P^2 := \mathrm{Var}_P(\xi_P \mid \widehat{f}), \tag{3.12}$$

and an analogous version of (3.11) without a subscript i. Further, write

$$\mathcal{E}_{P,1} := \frac{1}{n}\sum_{i=1}^{n}\{m_P(Z_i) - \widehat{m}(Z_i)\}^2, \qquad \mathcal{E}_{P,2} := \frac{1}{n\sigma_P^2}\sum_{i=1}^{n}\{m_{P,\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i)\}^2. \tag{3.13}$$

The second MSPE $\mathcal{E}_{P,2}$ in the display above is normalised by the variance of the errors $\xi_{P,i}$ featuring in the corresponding regression. Under the null, we expect this variance to be small as $\widehat{f}$ is then estimating a zero function, and consequently $\mathcal{E}_{P,2}$ may be inflated. On the other hand, as $\widehat{f}$ is small, we can expect that the unnormalised MSPE is particularly small. For example, writing $\mathcal{P}$ for the simple null linear model considered in (3.7), we would have

$$\frac{1}{n}\sum_{i=1}^{n}\{m_{P,\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i)\}^2 = O_{\mathcal{P}}(n^{-2}) \qquad \text{and} \qquad 1/\sigma_P^2 = O_{\mathcal{P}}(n),$$

giving $\mathcal{E}_{P,2} = O_{\mathcal{P}}(n^{-1})$. We are now in a position to present our results on type I error and power.

3.4.1 Type I error control

We consider the following assumption regarding general Type I error control.

Assumption 3.3. Consider a class of null distributions $\mathcal{P}_0$ of (X, Y, Z) on $\mathcal{X} \times \mathbb{R} \times \mathcal{Z}$ with $\mathbb{E}_P(Y \mid X, Z) = \mathbb{E}_P(Y \mid Z)$ for which there exists $c > 0$ such that $\inf_{P \in \mathcal{P}_0} \mathbb{E}_P(\varepsilon_P^2 \mid X, Z) \geq c$, and the following hold.

(a) $\sup_{P \in \mathcal{P}_0} \mathbb{P}_P(\sigma_P^2 = 0) = o(1)$.

(b) The product of the MSPEs satisfies $\mathcal{E}_{P,1}\mathcal{E}_{P,2} = o_{\mathcal{P}_0}(n^{-1})$.

(c) The weighted MSPEs scaled by σ_P^2 satisfy

$$\frac{1}{n\sigma_P^2} \sum_{i=1}^{n} \{m_P(Z_i) - \widehat{m}(Z_i)\}^2 \xi_{P,i}^2 = o_{\mathcal{P}_0}(1)$$
$$\frac{1}{n\sigma_P^2} \sum_{i=1}^{n} \{m_{P,\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i)\}^2 \varepsilon_{P,i}^2 = o_{\mathcal{P}_0}(1).$$

(d) There exists $\delta \in (0, 2]$ such that $\mathbb{E}_P(|\varepsilon_P \xi_P|^{2+\delta} \mid \widehat{f})/\sigma_P^{2+\delta} = o_{\mathcal{P}_0}(n^{\delta/2})$.

Part (a) of Assumption 3.3 asks that the test statistic should be asymptotically non-degenerate. As discussed in Section 3.2, degeneracy causes difficulties for establishing the asymptotic normality, but does not preclude Type I error control. For instance, the variance σ_P^2 is zero if the direction estimate $\widehat{f}$ is constant in X. This situation, which is in favour of the null, can be checked empirically, and the Type I error can still be controlled if one does not reject the null whenever this degenerate solution occurs.

Part (b) should be regarded as the primary restriction on $\mathcal{P}_0$, and along with (c), relates directly to the performance of the user-chosen regression methods involved in the construction of the PCM. As alluded to, in a simple linear model setting, we can expect $\mathcal{E}_{P,1}\mathcal{E}_{P,2} = O_{\mathcal{P}_0}(n^{-2})$, which certainly satisfies the condition. The rate requirement on the product of MSPEs is however sufficiently slow to accommodate nonparametric models; see Section 3.5. We note that the deterministic condition $\sup_{P \in \mathcal{P}_0}\{\mathbb{E}(\mathcal{E}_{P,1})\mathbb{E}(\mathcal{E}_{P,2})\} = o(n^{-1})$ is sufficient to guarantee part (b), as can be verified via Markov's inequality and the Cauchy–Schwarz inequality. If in addition there exists $C > 0$ such that $\mathrm{Var}(\xi_P \mid Z, \widehat{f}) \leq C\sigma_P^2$ and $\mathrm{Var}(\varepsilon_P) \leq C$, then (c) is guaranteed when $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ satisfy the simple consistency properties $\mathbb{E}(\mathcal{E}_{P,1})$, $\mathbb{E}(\mathcal{E}_{P,2}) = o(1)$. Part (d) is a conditional Lyapunov condition, and is used to apply the central limit theorem for triangular arrays.

Theorem 3.1 (Asymptotic normality under the null of a general procedure)**.** *Suppose that Assumption 3.3 holds over a class of null distributions $\mathcal{P}_0$. Then*

$$\sup_{P \in \mathcal{P}_0} \sup_{t \in \mathbb{R}} \left| \mathbb{P}_P(T \leq t) - \Phi(t) \right| \to 0.$$

As a consequence, the test $\mathbb{1}_{\{T > z_{1-\alpha}\}}$ has uniform asymptotic size α over $\mathcal{P}_0$.

The proof of Theorem 3.1 can be found in Appendix 3.8.4, which formalises a brief explanation of normality laid down in Section 3.2.1. The above result indicates that the asymptotic normality of T (hence the validity of the PCM test) is largely determined by the

predictive performance of regression models used in construction of the test statistic. This is similar in spirit to the prior work of Lundborg et al. (2021a); Shah and Peters (2020), but our proposal involves an additional regression step for the projection. As we discussed earlier, this extra step is critical to obtaining significant power against broader classes of alternatives, yet resulting in more delicate conditions for Type I error control.

As part of the proof of Theorem 3.1, we impose a condition that is trivially satisfied when $\widehat{m}$ is formed on an auxiliary sample. The conclusion of Theorem 3.1 for the practical version of our test where $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ are formed on $\mathcal{D}_1$ follows under the same conditions when we further assume that $X \perp\!\!\!\perp Y \,|\, Z$. It also follows immediately if $\widehat{m}$ is a linear smoother. See Proposition 3.6 in Appendix 3.9.2 together with the proof of Theorem 3.1 for more details.

3.4.2 Power properties

When studying the power properties of our test, we restrict attention to a subset of alternatives that are separated from null distributions characterised by Assumption 3.4 below.

Assumption 3.4. Given positive sequence $(\epsilon_n)_{n\in\mathbb{N}}$, let $(\mathcal{P}_1(\epsilon_n))_{n\in\mathbb{N}}$ be a sequence of collections of alternative distributions such that

$$\inf_{P\in\mathcal{P}_1(\epsilon_n)} \tau_P \geq \epsilon_n.$$

Further, suppose there exists $C > 0$ with

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_1(\epsilon_n)} \max\{\mathrm{Var}_P(Y \,|\, X, Z), h_P(X, Z)\} \leq C,$$

and that the following conditions are satisfied.

(a) There exists $\beta_1 > 0$ such that $\mathcal{E}_{P,1} = O_{\mathcal{P}_1(\epsilon_n)}(n^{-\beta_1})$.

(b) There exists $\beta_2 > 0$ such that $\mathcal{E}_{P,2} = O_{\mathcal{P}_1(\epsilon_n)}(n^{-\beta_2})$

(c) There exists $\rho > 0$ such that

$$\sup_{P\in\mathcal{P}_1(\epsilon_n)} \mathbb{P}_P\Big(\mathrm{Corr}_P\big(h_P(X, Z), \xi_P \,|\, \widehat{f}\big) \leq \rho\Big) = o(1).$$

In addition to the rate requirements on MSPEs in (a) and (b), condition (c) requires ξ_P, the population residuals from regressing our estimated projection $\widehat{f}$ onto Z, to be positively correlated with h_P with high probability. To interpret (c), it is helpful to consider a stronger version with $\widehat{f}(X, Z)$ replacing ξ_P (note that $\mathbb{E}_P\big(h_P(X, Z)\xi_P \,|\, \widehat{f}\big) = \mathbb{E}_P\big(h_P(X, Z)\widehat{f}(X, Z) \,|\, \widehat{f}\big)$ and $\mathbb{E}(\xi_P^2 \,|\, \widehat{f}) = \mathbb{E}_P[\mathrm{Var}_P(\widehat{f}(X, Z) \,|\, Z, \widehat{f}) \,|\, \widehat{f}] \leq \mathbb{E}(\widehat{f}(X, Z)^2 \,|\, \widehat{f})$). This latter condition permits $\widehat{f}$ to be an inconsistent estimator of the true $f_P(X, Z) = h_P(X, Z)/\mathrm{Var}_P(Y \,|\, X, Z)$.

The flexibility of this assumption relies on using regression method $\widehat{m}_{\widehat{f}}$ being scale equivariant in the sense that

$$\widehat{m}_{a\cdot\widehat{f}}(Z) = a \cdot \widehat{m}_{\widehat{f}}(Z) \tag{3.14}$$

for all $a > 0$; such a property however would be satisfied by almost all regression methods.

We can now state the main result of this subsection.

Theorem 3.2. *Assume that $\widehat{m}_{\widehat{f}}$ is scale equivariant, that is, satisfying (3.14) and consider the sequence of classes of distributions in Assumption 3.4, where*

$$\epsilon_n \cdot n^{\min\{1,\beta_1+\beta_2\}} \to \infty. \tag{3.15}$$

Then for any $\alpha \in (0,1)$,

$$\inf_{P \in \mathcal{P}_1(\epsilon_n)} \mathbb{P}_P(T > z_{1-\alpha}) \to 1.$$

Theorem 3.2 shows that if $\beta_1 + \beta_2 \geq 1$, as may be expected if m and $m_{\widehat{f}}$ are sufficiently smooth, the rate ϵ_n is going to be driven by Assumption 3.4 (c).

It is possible to derive a version of Theorem 3.2 for the test as described in Algorithm 2 which does not employ the additional sample splitting we are considering here. The only change is that (3.15) becomes $\epsilon_n \cdot n^{\min\{1,\beta_1,\beta_2\}} \to \infty$; however we believe the version of Theorem 3.2 above is more in line with the behaviour to be expected in practice, and empirically we find the version of the test in Algorithm 2 to provide better discrimination between null and alternatives in finite samples.

3.5 Series estimators

Following the theory in the previous section for a general regression method, we will now provide more concrete results for a specific class, namely spline estimators. In particular, our interest is to identify conditions under which our test is uniformly asymptotically valid and attains near-optimal power in a nonparametric setting. A formal power analysis, however, is complicated by the fact that $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ are computed on the same subsample as our test statistic. We therefore consider an alternative test statistic that leverages ideas from the literature on cross-fitting (Newey and Robins, 2018), a tool for reducing estimation bias for methods based on sample splitting. This additional sample splitting is a key component in obtaining a fast separation rate in our power analysis, while retaining uniform asymptotic Type I error control. Nevertheless, extra sample splitting may sacrifice finite-sample performance, and we therefore consider this variant mainly for theoretical purposes. Throughout this section we assume that $(X, Z) \in [0,1]^{d_X} \times [0,1]^{d_Z}$ and set $d := d_X + d_Z$.

We start by describing the test statistic as constructed using spline regression estimators with additional sample splitting. We will require two additional independent samples of size n, so that we have $\mathcal{D}_1, \ldots, \mathcal{D}_4$ in total. In Appendix 3.10, we give a self-contained description of spline spaces and their tensor product B-spline bases, containing all the results that we require for our analysis. In particular, given a spline order $r \in \mathbb{N}$ (i.e. degree $r-1$) and $N \in \mathbb{N}_0$ equispaced interior knots in each dimension, we denote by $\mathcal{S}_{r,N}^{d_Z}$ the corresponding spline space on $[0,1]^{d_Z}$, and by $\boldsymbol{\phi}^Z$ its d_Z-tensor B-spline basis, which consists of $K_Z := (N+r)^{d_Z}$ basis functions. Writing $\mathcal{S}_{r,N}^{d_X}$ for the corresponding spline space on $[0,1]^{d_X}$ with d_X-tensor B-spline basis $\boldsymbol{\phi}^X$, having $K_X := (N+r)^{d_X}$ basis functions, we can define the d-tensor product basis

$\boldsymbol{\phi}(x,z) := \boldsymbol{\phi}^X(x) \otimes \boldsymbol{\phi}^Z(z)$ for $\mathcal{S}^d_{r,N}$, where $\boldsymbol{u} \otimes \boldsymbol{v} := \mathrm{vec}(\boldsymbol{u}\boldsymbol{v}^\top)$, having $K_{XZ} := K_X K_Z$ basis functions. Further, we let $\boldsymbol{\psi}$ denote the tensor product B-spline basis for $\mathcal{S}^{d_Z}_{2r-1,N}$, and write $\widetilde{K}_Z := (N+2r-1)^{d_Z}$; the higher order of the spline basis functions that make up $\boldsymbol{\psi}$ affords better approximation properties that turn out to be useful for our theory.

The following description of the test statistic follows Algorithm 2 except that we fit $\widehat{m}_{\widehat{f}}$ on $\mathcal{D}_3$, $\widehat{m}$ on $\mathcal{D}_4$ and set $\widehat{v} \equiv 1$ for simplicity. We will also omit discussion of the sign correction step (Algorithm 2 1(iv)), since $\hat{\rho}$ is always nonnegative for the estimators considered below. We first regress Y onto $\boldsymbol{\phi}(X,Z)$ using ordinary least squares (OLS) on $\mathcal{D}_2$, yielding an estimator $\widehat{\boldsymbol{\beta}}_{XZ} \in \mathbb{R}^{K_{XZ}}$, and set $\widehat{g}(x,z) := \widehat{\boldsymbol{\beta}}_{XZ}^\top \boldsymbol{\phi}(x,z)$. Next, we regress $\widehat{g}(X,Z)$ onto $\boldsymbol{\phi}(X,Z)$ using OLS on $\mathcal{D}_2$ again, to obtain an estimator $\widehat{\boldsymbol{\beta}}_Z \in \mathbb{R}^{K_Z}$, and set $\widetilde{m}(z) := \widehat{\boldsymbol{\beta}}_Z^\top \boldsymbol{\phi}^Z(z)$. Note that this is equivalent to regressing Y onto $\boldsymbol{\phi}^Z(Z)$. We then define the projection $\widehat{f}(x,z) := \widehat{g}(x,z) - \widetilde{m}(z)$. Using the fact that $\boldsymbol{\phi}^X$ forms a partition of unity (Proposition 3.9(a)), it follows that if we write $\widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}_{XZ} - \mathbf{1} \otimes \widehat{\boldsymbol{\beta}}_Z$, where $\mathbf{1} \in \mathbb{R}^{K_X}$ denotes a vector of ones, then $\widehat{f}(x,z) = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\phi}(x,z)$.

To estimate $m_{\widehat{f}}$, we regress $\widehat{f}(X,Z)$ onto $\boldsymbol{\psi}(Z)$ on $\mathcal{D}_3$ using OLS, yielding an estimator $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^{\widetilde{K}_Z}$, and set $\widehat{m}_{\widehat{f}}(z) := \widehat{\boldsymbol{\theta}}^\top \boldsymbol{\psi}(z)$. Similarly, we estimate m by regressing Y onto $\boldsymbol{\psi}(Z)$ on $\mathcal{D}_4$ using OLS, yielding an estimator $\widehat{\boldsymbol{\gamma}} \in \mathbb{R}^{\widetilde{K}_Z}$, and set $\widehat{m}(z) := \widehat{\boldsymbol{\gamma}}^\top \boldsymbol{\psi}(z)$. Given $\widehat{f}$, $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ as defined above, we compute the test statistic as in Algorithm 2 (on $\mathcal{D}_1$) and denote it by T_{Spline}. In Theorem 3.3 in Section 3.5.1 below, we demonstrate that T_{Spline} enjoys uniform asymptotic Type I error control under appropriate regularity conditions, while Theorem 3.4 and Proposition 3.4 in Section 3.5.2 reveal that T_{Spline} can achieve the optimal testing rate for this problem.

3.5.1 Type I error control

We start by stating our main distributional assumptions, which rely on the definitions of Hölder spaces $\mathcal{H}^d_s$ and Hölder norms $\|\cdot\|_{\mathcal{H}_s}$ given in Definition 3.3.

Assumption 3.5. Let $\mathcal{P}$ be a class of distributions of (X,Y,Z) on $[0,1]^{d_X} \times \mathbb{R} \times [0,1]^{d_Z}$, and for $P \in \mathcal{P}$, let $m_P(z) := \mathbb{E}_P(Y \,|\, Z = z)$, $\varepsilon_P := Y - m_P(Z)$ and $g_P(x,z) := \mathbb{E}_P(Y \,|\, X = x, Z = z)$. Assume that there exist $C \geq 1$ and $c \in (0,1]$ with the following properties:

(a) For each $P \in \mathcal{P}$, we have $\mathbb{E}_P(\varepsilon_P^2 \,|\, X, Z) \geq c$ and there exists $\delta \in (0,2]$ such that $\mathbb{E}_P(|\varepsilon_P|^{2+\delta} \,|\, X, Z) \leq C$.

(b) For each $P \in \mathcal{P}$, the marginal distribution of (X,Z) is absolutely continuous with respect to Lebesgue measure on $[0,1]^d$, with density p_P satisfying $\sup_{(x,z)\in[0,1]^d} p_P(x,z) \leq C$ and $\inf_{(x,z)\in[0,1]^d} p_P(x,z) \geq c$.

(c) Let $s \in (0,r]$ and let $p_{X|Z,P}(\cdot \,|\, z)$ denote the conditional density of X given $Z = z$. Assume that for all $P \in \mathcal{P}$, we have $p_{X|Z,P}(x \,|\, \cdot) \in \mathcal{H}^{d_Z}_s$ for all $x \in [0,1]^{d_X}$, and that $m_P \in \mathcal{H}^{d_Z}_s$ and $g_P \in \mathcal{H}^d_s$, with

$$\max\Big\{ \sup_{x\in[0,1]^{d_X}} \|p_{X|Z,P}(x,\cdot)\|_{\mathcal{H}_s}, \|m_P\|_{\mathcal{H}_s}, \|g_P\|_{\mathcal{H}_s} \Big\} \leq C.$$

Assumption 3.5 is closely related to other assumptions commonly used in spline regression (e.g. Belloni et al., 2015; Ichimura and Newey, 2015; Newey and Robins, 2018). We consider the case where all nuisance functions have the same smoothness for convenience only. In order to state our Type I error control result for T_{Spline}, it will be convenient to define the projection $\boldsymbol{\Pi} : \mathbb{R}^{K_{XZ}} \to \mathbb{R}^{K_{XZ}}$ by $\boldsymbol{\Pi}(\boldsymbol{x}) \equiv \boldsymbol{\Pi}(x_1, \ldots, x_{K_{XZ}}) := \boldsymbol{x} - \mathbf{1} \otimes \bar{\boldsymbol{x}}$, with $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_{K_Z})$ given by $\bar{x}_k := K_X^{-1} \sum_{\ell=1}^{K_X} x_{(k-1)K_X+\ell}$ for $k \in [K_Z]$.

Theorem 3.3 (Asymptotic normality of T_{Spline})**.** *Suppose that Assumption 3.5 holds for a class of null distributions $\mathcal{P}_0$, i.e. a class of distributions that also satisfies $\mathbb{E}_P(Y \mid X, Z) = \mathbb{E}_P(Y \mid Z)$ for every $P \in \mathcal{P}_0$. Assume that $\sup_{P \in \mathcal{P}_0} \mathbb{P}_P(\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty = 0) = o(1)$ and that $\boldsymbol{\Lambda}_P := \mathbb{E}_P\{\mathrm{Cov}_P(\boldsymbol{\phi}(X, Z) \mid Z)\}$ satisfies*

$$\tilde{\lambda}_{\min}(\boldsymbol{\Lambda}_P) := \min_{\boldsymbol{x} \in \mathbb{R}^{K_{XZ}} : \boldsymbol{\Pi}\boldsymbol{x} = \boldsymbol{x}, \|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^\top \boldsymbol{\Lambda}_P \boldsymbol{x} \geq \frac{c}{K_{XZ}} \tag{3.16}$$

where $c \in (0, 1]$ is taken from Assumption 3.5. Finally, suppose that

$$n K_{XZ} \left\{ \widetilde{K}_Z^{-2s/d_Z} + \frac{\widetilde{K}_Z}{n} \right\}^2 \to 0 \tag{3.17}$$

and

$$\frac{K_{XZ}^{1+2/\delta}}{n} \to 0 \tag{3.18}$$

where δ is taken from Assumption 3.5. Then

$$\sup_{P \in \mathcal{P}_0} \sup_{t \in \mathbb{R}} |\mathbb{P}_P(T_{\text{Spline}} \leq t) - \Phi(t)| \to 0.$$

The proof of Theorem 3.3 amounts to the verification of Assumption 3.3, which then allows us to apply our general Type I error control result, namely Theorem 3.1. In addition to Assumption 3.5, Theorem 3.3 imposes several additional conditions. The assumption that $\sup_{P \in \mathcal{P}_0} \mathbb{P}_P(\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty = 0) = o(1)$ simply avoids degeneracy of the test statistic and is used to show that Assumption 3.3(a) is satisfied. If this condition is not satisfied, then since we defined $0/0 := 0$ in the definition of our test statistic, it can be shown that

$$\liminf_{n \to \infty} \inf_{P \in \mathcal{P}_0} \mathbb{P}_P(|T_{\text{Spline}}| \leq t) \geq \Phi(t) - \Phi(-t)$$

for all $t \geq 0$ (i.e. $|T_{\text{Spline}}|$ is asymptotically stochastically dominated by the absolute value of a standard Gaussian random variable), so the test retains uniform asymptotic Type I error control provided that $\alpha \leq 1/2$.

Condition (3.16) can be regarded as a restricted minimum eigenvalue condition; for $\boldsymbol{x} \in \mathbb{R}^{K_{XZ}}$ with $\boldsymbol{\Pi}\boldsymbol{x} = 0$, we have that $\boldsymbol{x}^\top \boldsymbol{\Lambda}_P \boldsymbol{x} = 0$, but it turns out that we are able to restrict attention to the orthogonal complement of this subspace. Motivation for the form of this condition is provided by the fact that, writing $\boldsymbol{\Sigma}_P := \mathbb{E}_P(\boldsymbol{\phi}(X, Z)\boldsymbol{\phi}(X, Z)^\top) \in \mathbb{R}^{K_{XZ} \times K_{XZ}}$, we have

$$\tilde{\lambda}_{\min}(\boldsymbol{\Lambda}_P) \leq \tilde{\lambda}_{\min}(\boldsymbol{\Sigma}_P) \leq \lambda_{\max}(\boldsymbol{\Sigma}_P) \leq C 2^d K_{XZ}^{-1}$$

by Proposition 3.9(d). Moreover, Lemma 3.11 in Section 3.9.2 of the appendix shows that the assumption holds when X and Z are independent.

Condition (3.17) is used to show that parts (b) and (c) of Assumption 3.3 are satisfied while condition (3.18) is used to show that part (d) of the assumption is satisfied. These conditions control the interplay between on the growth rate of the number of basis functions, the smoothness s of the regression functions and conditional densities and δ. When choosing the knot spacing to minimise the mean-squared error of the involved regressions, we would choose $\widetilde{K}_Z$ and K_Z of order $n^{d_Z/(2s+d_Z)}$ and K_X of order $n^{d_X/(2s+d_Z)}$. Thus, for (3.17) to hold, we need $s > d_Z + d_X/2$ and for (3.18) to hold, we need $\delta > 2(d_X + d_Z)/(2s - d_X)$. Both of conditions could be weakened, at the expense of additional notational complexity, by choosing different knot spacings N_X and N_Z for the d_X- and d_Z-tensor B-spline bases $\boldsymbol{\phi}^X$ and $\boldsymbol{\phi}^Z$ for our spline spaces $\mathcal{S}^{d_X}_{r,N_X}$ and $\mathcal{S}^{d_Z}_{r,N_Z}$. Indeed, by taking N_X, and hence K_X, to be of constant order, while retaining the original choices of K_Z and $\widetilde{K}_Z$, we see that (3.17) holds when $s > d_Z$ and (3.18) holds when $\delta > d_Z/s$ (so it would suffice for Assumption 3.5(a) to hold with $\delta = 1$, provided again that $s > d_Z$).

3.5.2 Power and minimax lower bound

As mentioned at the beginning of this section, we employ additional sample splitting in the construction of T_{Spline}. This turns out to be key in demonstrating the optimality of our test. To provide insight into the benefits of sample splitting in this context, consider two generic spline estimators $\widehat{g}_1$ and $\widehat{g}_2$ of unknown functions g_1 and g_2, respectively. Suppose that we would like to choose $\widehat{g}_1$ and $\widehat{g}_2$ to minimise the empirical cross-product error

$$\widehat{\theta}_{\text{cross}} := \frac{1}{n}\sum_{i=1}^{n}\{\widehat{g}_1(Z_i) - g_1(Z_i)\}\{\widehat{g}_2(Z_i) - g_2(Z_i)\}. \tag{3.19}$$

A naive way of approaching this problem is to construct $\widehat{g}_1$ and $\widehat{g}_2$ on the same dataset and to choose the number of spline functions to minimise the mean-squared error of each of $\widehat{g}_1$ and $\widehat{g}_2$. The Cauchy–Schwarz inequality then guarantees that the cross-product error is small as long as the mean-squared errors are small. However, this indirect approach returns a potentially suboptimal rate of convergence due to its "own observation" bias, which arises from using the same datasets to form $\widehat{g}_1$ and $\widehat{g}_2$. When employing auxiliary samples to construct $\widehat{g}_1$ and $\widehat{g}_2$ we can eliminate this bias — an idea originally proposed by Newey and Robins (2018). Thus, a more refined analysis of terms like $\widehat{\theta}_{\text{cross}}$ that does not directly employ the Cauchy–Schwarz inequality can result in faster convergence rates; see for instance Proposition 3.15 in the appendix. Our main result in this section is as follows:

Theorem 3.4. *Let $\mathcal{P}$ be a class of distributions satisfying Assumption 3.5, and let $\mathcal{P}_1(\epsilon_n) := \{P \in \mathcal{P} : \tau_P \geq \epsilon_n\}$, where*

$$\epsilon_n \cdot n^{\frac{4s}{4s+d}} \to \infty. \tag{3.20}$$

Further, assume that the tuning parameters are chosen such that $K_X \asymp n^{\frac{2d_X}{4s+d}}$ *and* $K_Z \asymp \widetilde{K}_Z \asymp n^{\frac{2d_Z}{4s+d}}$ *and that* $r \geq s \geq 3d/4$. *Then*

$$\inf_{P \in \mathcal{P}_1(\epsilon_n)} \mathbb{P}_P(T_{\text{Spline}} > z_{1-\alpha}) \to 1.$$

Theorem 3.4 reveals that the test based on T_{Spline} has uniform asymptotic power 1 over a class of alternatives that are sufficiently separated from the null, as defined by $\mathcal{P}_1(\epsilon_n)$.

We remark that in Theorem 3.4, we have operated in the context of a known smoothness parameter s for theoretical purposes. It is possible to construct more involved tests that adapt to unknown smoothness levels following the strategy of Lepskiĭ (1991) and Ingster (2000), but we do not pursue this direction further.

The separation rate (3.20) cannot be improved further from a minimax perspective, as illustrated by the following lower bound result.

Proposition 3.4. *Consider a class of distributions, denoted by* $\mathcal{P}$, *that satisfy Assumption 3.5, and let* $\mathcal{P}_1(\epsilon_n) := \{P \in \mathcal{P} : \tau_P \geq \epsilon_n\}$. *Then, for a fixed level* $\alpha \in (0, 1/2)$, *there exists* $c > 0$ *such that if* $\limsup_{n\to\infty} \epsilon_n \cdot n^{\frac{4s}{4s+d}} < c$, *then any test* ϕ *having uniform asymptotic size at most* α *satisfies*

$$\limsup_{n\to\infty} \inf_{P \in \mathcal{P}_1(\epsilon_n)} \mathbb{P}_P(\phi = 1) \leq \alpha + 1/2.$$

Proposition 3.4 complements Theorem 3.4 by showing that τ_P is a small constant multiple of $n^{-\frac{4s}{4s+d}}$, no test can achieve uniform consistency under Hölder smoothness. The proof of Proposition 3.4, which can be found in Appendix 3.8.8, follows a fairly standard argument (e.g. Arias-Castro et al., 2018; Ingster, 1987) that bounds the χ^2-divergence from a fixed null distribution to a mixture of distributions in the alternative class $\mathcal{P}_1(\epsilon_n)$.

3.6 Numerical experiments

In this section, we present the results of several simulation experiments that investigate the empirical performances of both the recommended multiple sample splitting version of the PCM (see Algorithm 3) with $B = 6$ splits, denoted by `pcm`, and the single split version (see Algorithm 2) denoted by `pcm_ss`. We compare our tests to various conditional (mean) independence tests in the literature listed below.

`gam` The test based on the default p-value for a smooth when fitting a generalised additive model (GAM) using the `mgcv`-package in R (Wood, 2013, 2017).

`williamson` The test resulting from applying the approach described in Williamson et al. (2021) and employing sample splitting as implemented in the `vim` function from the `vimp`-package in R (Williamson et al., 2022).

`kci` The *kernel conditional independence test* (Zhang et al., 2011) as implemented in the `KCI` function of the `CondIndTests` R package (Heinze-Deml et al., 2018); we use the Bayesian hyperparameter tuning option for sample sizes of at most 500.

`gcm` The *Generalised Covariance Measure* (GCM) as described in Shah and Peters (2020).

`wgcm.fix` The 'fixed weight function' variant of the *Weighted Generalised Covariance Measure* (wGCM) (Scheidegger et al., 2021) as implemented in the `wgcm.fix` function of the `weightedGCM` R package; we use `weight.num` $= 7$ as in the simulations of the original paper.

`wgcm.est` The 'estimated weight function' variant of the wGCM as implemented in the `wgcm.est` function of the `weightedGCM` R-package.

In all of our numerical simulations, rejection rates were estimated based on 100 repetitions.

3.6.1 Additive models

We first investigate Type I error control and power in setting where both $\mathbb{E}(Y \mid Z)$ and $\mathbb{E}(X \mid Z)$ are additive functions, and $Z \sim N_7(0, \boldsymbol{I})$. For the methods, including the PCM, requiring choices of regression procedures, we use an additive model fitted using `mgcv`. We consider null settings consisting of $n \in \{250, 500, 1000\}$ independent and identically distributed copies of (X, Y, Z) where

$$X = \sin(2\pi Z_1) + 0.1\xi \quad \text{and} \quad Y = \sin(2\pi Z_1) + \varepsilon.$$

and errors ε and ξ are independent $N(0,1)$ random variables, independent of Z. Such a setup is challenging for Type I error control as X and Y are highly correlated yet are conditionally independent given Z. Indeed we see from the left panel of Figure 3.2 that several of the tests are anti-conservative, most notably `kci` and `gam`, which we omit from further comparisons as their power properties would be hard to interpret given the high rejection rates under the null. Both the `williamson` and `wgcm.est` tests are also somewhat anti-conservative, but considerably less so. In contrast, the `pcm` is conservative here. This is to be expected as the calibration following the multiple sample splits involved in its construction (Section 3.2.2) is typically conservative; the single split version `pcm_ss` appears to have rejection rates close to the 5% mark as suggested by our theory.

We investigate the power properties of the PCM in the following settings, where as before, ε and ξ are independent and independent of Z, and moreover $\varepsilon \sim N(0,1)$.

1. $\xi \sim N(0,1)$, $X = \sin(2\pi Z_1) + \xi$ and $Y = \sin(2\pi Z_1) + 0.2X^2 + \varepsilon$.
2. $\xi + 1 \sim \text{Exp}(1)$, $X = \sin(2\pi Z_1) - \sin(2\pi Z_1)\xi$ and $Y = \sin(2\pi Z_1) + 0.4X^2 + \varepsilon$.
3. $\xi \sim N(0,1)$, $X = \sin(2\pi Z_1) + \xi$ and $Y = \sin(2\pi Z_1) + 0.4X^2 Z_2 + \varepsilon$.

The settings are chosen such that in setting 1: $\mathbb{E}(\text{Cov}(X, Y \mid Z)) = 0$ but $\text{Cov}(X, Y \mid Z) \neq 0$, in setting 2: $\text{Cov}(X, Y \mid Z) = 0$ but $\tau \neq 0$ and in setting 3: there is only an interaction effect. We cannot expect this interaction effect to be picked up by methods that fit additive models,

Fig. 3.2 Rejection rates in the various settings considered in Section 3.6.1 for nominal 5%-level tests. Note that Alternative 3 has only an interaction effect, so we cannot expect methods that fit additive models to have power.

but we nevertheless include this setting to emphasise the fact that the success of the PCM and related methods is contingent on an appropriate choice of regression method; see also Section 3.6.2.

From the right-hand panels of Figure 3.2, we see that the `pcm` and `williamson` exhibit good power in settings 1 and 2, with the latter also rejecting in setting 3; however the rejection rates for `williamson` should be interpreted carefully given the miscalibration in the null settings. `wgcm.est` also shows appreciable power in setting 1, though as expected has little power in setting 2 where $\mathrm{Cov}(X, Y \mid Z) = 0$. For the reasons explained above, the PCM has no power in setting 3; in the next section we investigate the performance of the PCM when used in conjunction with a regression method capable of fitting to such regression functions.

3.6.2 Non-additive models

In this section, we consider settings where the regression functions are non-additive and involve complex interactions. We use random forests (Breiman, 2001) implemented in the `ranger` R package (Wright and Ziegler, 2017) as our regression procedure for the methods considered.

We consider null settings consisting of $n \in \{10^4, 2 \cdot 10^4, 4 \cdot 10^4\}$ independent and identically distributed copies of (X, Y, Z) where $Z \sim N_7(0, \boldsymbol{I})$ as before,

$$X = \sin(2\pi Z_1)(1 + Z_3) + \xi \quad \text{and} \quad \sin(2\pi Z_1)(1 + Z_3) + v(X)\varepsilon$$

with ε and ξ independent $N(0,1)$ random variables independent of Z, and $v(X) := 0.5 + \mathbb{1}_{\{X>0\}}$ giving heteroscedastic errors for the Y regression model. The larger sample sizes considered here reflect the difficulty of estimating the more complicated regression functions in these examples. Note that here we do not have $X \perp\!\!\!\perp Y \mid Z$, but the conditional mean independence $\mathbb{E}(Y \mid X, Z) = \mathbb{E}(Y \mid Z)$ does hold. The results are presented in Figure 3.3. We see that the multiple sample splitting version of the PCM maintains Type I error control, and is in fact slightly conservative. All other approaches considered appear to be anti-conservative to varying degrees: the `williamson` approach is most clearly miscalibrated here, and we omit it from our alternative settings described below; `wgcm.est` and `gcm` are also fairly anti-conservative here but the rejection rates appear to be improving for increasing n.

We consider the following alternative settings, where as in Section 3.6.1, setting 2 has $\mathrm{Cov}(X, Y \mid Z) = 0$, and setting 3 involves a pure interaction effect:

1. $\xi \sim N(0,1)$, $X = \sin(2\pi Z_1)(1 + Z_3) + \xi$ and $Y = \sin(2\pi Z_1)(1 + Z_3) + 0.04X^2 + v(X)\varepsilon$;
2. $\xi \sim \mathrm{Exp}(1)$, $X = \sin(2\pi Z_1)(1 + Z_3) - \sin(2\pi Z_1)(\xi - 1)$ and $Y = \sin(2\pi Z_1)(1 + Z_3) + 0.04X^2 + v(X)\varepsilon$;
3. $\xi \sim N(0,1)$, $X = \sin(2\pi Z_1)(1 + Z_3) + \xi$ and $Y = \sin(2\pi Z_1)(1 + Z_3) + 0.04X^2 Z_2 + v(X)\varepsilon$.

Among the methods considered, here only the PCM appear to have good power across the settings considered. The `wgcm.est` has reasonable power in setting 1, though this should be interpreted with some care given that Type I error is not very well controlled in the null settings. However in setting 2, `wgcm.est` is powerless as expected.

3.7 Conclusion

In this work we have introduced a general test statistic we call the PCM for testing conditional mean independence that: (a) can leverage machine learning methods to yield provable uniform Type I error control across a class of null distributions where these methods have sufficiently good predictive ability; and (b) when used in conjunction with appropriate regression methods attains rate-optimal power in both the parametric setting of the linear model and fully nonparametric settings. We believe the PCM fills an important gap in the data analyst's range of existing tools, which are unable to simultaneously achieve these desiderata. However, our work also offers several avenues for further work, some of which we mention below.

Verifying the general assumptions for other regression methods We have verified Assumption 3.3 for linear regression in linear model settings, and nonparametric series estimators in fully nonparametric settings. It would be interesting to see for example for what class $\mathcal{P}$, the penalised regression splines of `mgcv` used in several of our numerical experiments, satisfy this condition. Similarly, it would be very interesting to ask the same question of random forests, which perform very well in our simulations; however this is likely to be challenging given the complex nature of the random forest procedure.

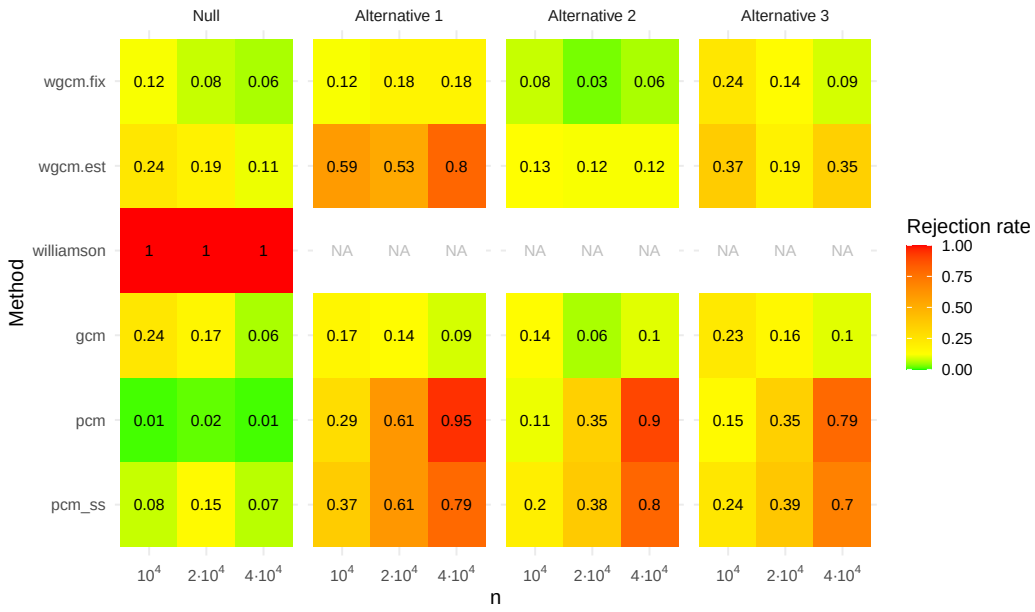


Fig. 3.3 Rejection rates in the various settings considered in Section 3.6.2 for nominal 5%-level tests.

Aggregation of test statistics from multiple sample splits Whilst our proposal (Algorithm 3) to average test statistics from multiple sample splits and compare this to a standard Gaussian quantile works well in practice, determining interpretable conditions under which this guarantees Type I error control is an open problem. Moreover, our numerical experiments suggest our proposal is somewhat conservative, and it is certainly of interest, both in our setting and other contexts involving multiple sample splitting, to develop an aggregation procedure that makes more efficient use of the sample to attain greater power.

Conditional independence testing Although the problem of testing conditional independence has been studied more intensively than that of testing conditional mean independence, there do not exist many practical methods that achieve both (a) and (b) for the former problem. One starting point for constructing a such a test may be the fact that the conditional independence null $Y \perp\!\!\!\perp X \,|\, Z$ may be viewed as the intersection of conditional mean independence nulls $\mathbb{E}(w(Y) \,|\, X, Z) = \mathbb{E}(w(Y) \,|\, Z)$ where function w ranges over all monotone functions, for example. It might therefore be interesting to investigate procedures that seek two 'projections': mappings $(X, Z) \to \widehat{f}(X, Z)$ and also $Y \mapsto \widehat{w}(Y)$, after which one may apply the GCM.

Confidence intervals We have focused on the problem of testing conditional mean independence, but the problem of deriving confidence intervals for a parameter such as τ that is 0 under our null is equally interesting. The pioneering work of Williamson et al. (2021) proposes an asymptotically optimal approach for this in the case where τ is bounded away from 0. It

would be interesting if the PCM could be used in conjunction with the proposal of Williamson et al. (2021) to extend the latter to yield confidence intervals with uniform coverage for all τ.

3.8 Proofs

In our proofs we often suppress the dependence of quantities on P for ease of notation.

3.8.1 Proof of Proposition 3.1

We will start by checking the assumptions of Lemma 3.10 for the regressions of Y on X and Z (of which $\widehat{\beta}$ is one component), Y on Z (yielding $\widehat{\boldsymbol{\theta}}$) and X on Z (yielding $\widehat{\boldsymbol{\eta}}$). Let $\delta' := \delta/2$.

Recalling $W := (X, Z)$, we see that *(i)* of Lemma 3.10 is satisfied for the Y on X and Z regression by our assumption on $\boldsymbol{\Sigma}^{XZ}$ and the fact that

$$\lambda_{\min}(\mathbb{E}(WW^\top\zeta^2)) \geq c\lambda_{\min}(\boldsymbol{\Sigma}^{XZ})$$

by our assumption that $\mathrm{Var}(Y \mid X, Z) \geq c$. *(ii)* is satisfied with $\delta = \delta'$ in that result by the Cauchy–Schwarz inequality and Jensen's inequality.

It is immediate from Assumption 3.1 that *(i)* of Lemma 3.10 is satisfied for the X on Z regression. To see that *(ii)* is satisfied with $\delta = \delta'$, we note that by the Cauchy–Schwarz inequality it suffices to check that $\mathbb{E}(|\xi|^{4+\delta})$ is bounded over $\mathcal{P}$. Letting $\boldsymbol{\Sigma} := \mathbb{E}(ZZ^\top)$, we have

$$\mathbb{E}(|\xi|^{4+\delta}) \leq 2^{3+\delta}\Big(\mathbb{E}(|X|^{4+\delta}) + \lambda_{\max}(\boldsymbol{\Sigma}^{-1})^{2+\delta/2}\|\mathbb{E}(XZ)\|_2^{4+\delta}\mathbb{E}(\|Z\|_2^{4+\delta})\Big)$$

which is bounded under Assumption 3.1.

To see that *(i)* of Lemma 3.10 is satisfied for the Y on Z regression, we note that

$$\mathbb{E}(ZZ^\top\{Y - \boldsymbol{\theta}^\top Z\}^2) = \mathbb{E}(ZZ^\top\{\zeta + X - \boldsymbol{\eta}^\top Z\}^2) = \mathbb{E}(ZZ^\top\zeta^2) + \mathbb{E}(ZZ^\top\{X - \boldsymbol{\eta}^\top Z\}^2),$$

and we have shown that both of these are lower bounded above. *(ii)* follows by similar arguments as those for the X on Z regression.

We now verify that Assumption 3.6 is satisfied. From the above it is immediate that (3.68) of Assumption 3.6 is satisfied and that

$$\sqrt{n_1}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 = O_{\mathcal{P}}(1) \quad \text{and} \quad \sqrt{n_1}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = O_{\mathcal{P}}(1). \tag{3.21}$$

We note that

$$\sup_{P\in\mathcal{P}} \mathbb{E}\left(\Big\|\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} Z_i\xi_i\Big\|_2^2\right) = \frac{1}{n_1}\sup_{P\in\mathcal{P}}\mathbb{E}(\|Z\xi\|_2^2) \to 0,$$

thus by Lemma 3.6 and (3.21) we have that (3.69) holds. Analogous arguments show that (3.70) is satisfied.

(3.21) shows that

$$\sqrt{n_1}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = O_{\mathcal{P}}(n_1^{-1/2}) = o_{\mathcal{P}}(1).$$

The proof of Lemma 3.10 and Assumption 3.1 yields that

$$\Big\|\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} Z_i Z_i^\top\Big\|_{\mathrm{op}} \leq \Big\|\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} Z_i Z_i^\top - \boldsymbol{\Sigma}\Big\|_{\mathrm{op}} + \|\boldsymbol{\Sigma}\|_{\mathrm{op}} = O_{\mathcal{P}}(1),$$

hence (3.71) is satisfied. The remaining conditions hold by similar arguments using the moment bounds established earlier, (3.21) and by applying Shah and Peters (2020, Lemma 19).

Arguments akin to those used to verify that Lemma 3.10 applies can be shown to verify the remaining conditions of Proposition 3.17 in Appendix 3.11. Therefore, the conclusion now follows, since,

$$\psi_{P,\alpha,n} = \Phi\Big(\frac{\kappa}{\sigma^2_{\beta_P}}\Big)\cdot\Phi\Big(z_\alpha + \frac{\kappa\sigma^2_{P,X|Z}}{\sigma_{P,XY|Z}}\Big) + \Phi\Big(-\frac{\kappa}{\sigma^2_{\beta_P}}\Big)\cdot\Phi\Big(z_\alpha - \frac{\kappa\sigma^2_{P,X|Z}}{\sigma_{P,XY|Z}}\Big),$$

which is increasing in κ.

3.8.2 Proof of Proposition 3.2

Throughout this proof we work on the event that at least one $u_{n,1},\dots,u_{n,n}$ is non-zero which is a set of uniform asymptotic probability 1 by Assumption 3.2(b). Let $\boldsymbol{Z} := (Z_1^\top,\dots,Z_n^\top)^\top$, $\boldsymbol{P} := \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$, $\boldsymbol{Y} := (Y_1,\dots,Y_n)^\top$, $\boldsymbol{\varepsilon} := (\varepsilon_1,\dots,\varepsilon_n)^\top$, $\widehat{\boldsymbol{f}} := \big(\widehat{f}(X_1,Z_1),\dots,\widehat{f}(X_n,Z_n)\big)^\top$ and $\boldsymbol{I}$ denote the $d\times d$ identity matrix. Since $\boldsymbol{P}$ is a matrix representing an orthogonal projection such that $\boldsymbol{Z}^\top(\boldsymbol{I}-\boldsymbol{P})$ is a zero vector, we have

$$\begin{aligned}\sum_{i=1}^n\{Y_i - \widehat{\boldsymbol{\gamma}}^\top Z_i\}\{\widehat{f}(X_i,Z_i) - \widehat{m}_{\widehat{f}}(Z_i)\} &= \widehat{\boldsymbol{f}}^\top(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} = \widehat{\boldsymbol{f}}^\top(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{\varepsilon}\\ &= \sum_{i=1}^n \varepsilon_i\{\widehat{f}(X_i,Z_i) - \widehat{m}_{\widehat{f}}(Z_i)\}.\end{aligned}$$

Based on the above identity, we have that

$$T_{\mathrm{OLS}} = \frac{\frac{1}{\sqrt{n}\nu}\sum_{i=1}^n \varepsilon_i u_{n,i}}{\sqrt{\frac{1}{n\nu^2}\sum_{i=1}^n (Y_i - \widehat{\boldsymbol{\gamma}}^\top Z_i)^2 u_{n,i}^2 - \big(\frac{1}{n\nu}\sum_{i=1}^n \varepsilon_i u_{n,i}\big)^2}},$$

where

$$\nu := \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\varepsilon_i^2 \mid X_i, Z_i) u_{n,i}^2} > 0.$$

Let $\mathcal{F}_n$ denote the σ-algebra generated by $\widehat{f}$ and $((X_i,Z_i))_{i=1}^n$. Form the triangular array

$$W_{n,i} := \frac{\varepsilon_i u_{n,i}}{\nu}$$

for $n\in\mathbb{N}$ and $i\in[n]$, and note that this satisfies the first three assumptions of Lemma 3.8, since $u_{n,i}$ is measurable with respect to $\mathcal{F}_n$. Finally, the fourth assumption of this lemma is

also satisfied, because

$$\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}(|W_{n,i}|^{2+\delta}\,|\,\mathcal{F}_n) = \frac{1}{n^{1+\delta/2}\nu^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}(|\varepsilon_i|^{2+\delta}\,|\,X_i,Z_i)|u_{n,i}|^{2+\delta}$$
$$\leq \frac{C}{c^{1+\delta/2}}\sum_{i=1}^{n}|\upsilon_{n,i}|^{2+\delta} \leq \frac{C}{c^{1+\delta/2}}\biggl(\sum_{i=1}^{n}|\upsilon_{n,i}|^2\biggr)\max_{i\in[n]}|\upsilon_{n,i}|^{\delta} = o_{\mathcal{P}_0}(1)$$

by Assumptions 3.2(a) and (b) and Lemmas 3.3 and 3.7. Lemma 3.8 thus yields that the numerator of T_{OLS} is uniformly asymptotically standard Gaussian.

For the denominator of T_{OLS}, the uniform version of Slutsky's theorem (Bengs and Holzmann, 2019, Theorem 6.3) yields that $\frac{1}{n\nu}\sum_{i=1}^{n}\varepsilon_i u_{n,i} = o_{\mathcal{P}_0}(1)$. For the first term in the denominator of T_{OLS}, we consider the decomposition

$$\frac{1}{n\nu^2}\sum_{i=1}^{n}(Y_i-\widehat{\boldsymbol{\gamma}}^{\top}Z_i)^2u_{n,i}^2 = \underbrace{\frac{1}{n\nu^2}\sum_{i=1}^{n}\varepsilon_i^2u_{n,i}^2}_{\mathrm{I}_n} + \underbrace{\frac{1}{n\nu^2}\sum_{i=1}^{n}\{(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^{\top}Z_i\}^2u_{n,i}^2}_{\mathrm{II}_n}$$
$$\underbrace{-\frac{2}{n\nu^2}\sum_{i=1}^{n}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^{\top}Z_i\varepsilon_iu_{n,i}^2}_{\mathrm{III}_n}.$$

Similarly to our previous argument, define the triangular array $\tilde{W}_{n,i} := W_{n,i}^2$ for $n\in\mathbb{N}$ and $i\in[n]$, and note that $(\tilde{W}_{n,i})_{n\in\mathbb{N},i\in[n]}$ satisfies the conditions of Lemma 3.9 with $\mu_n = 1$ in that result, so $\mathrm{I}_n = 1 + o_{\mathcal{P}_0}(1)$. Now, by Hölder's inequality,

$$|\mathrm{II}_n| \leq \frac{1}{n\nu^2}\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1^2\sum_{i=1}^{n}\|Z_i\|_\infty^2u_{n,i}^2 \leq \frac{1}{c}\max_{i\in[n]}\|Z_i\|_\infty^2\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1^2\sum_{i=1}^{n}\upsilon_{n,i}^2 = o_{\mathcal{P}}(1),$$

by Assumption 3.2 and Lemma 3.3. Finally, the Cauchy–Schwarz inequality yields that

$$|\mathrm{III}_n| \leq 2\sqrt{\frac{1}{n\nu^2}\sum_{i=1}^{n}\varepsilon_i^2u_{n,i}^2}\cdot\sqrt{\frac{1}{n\nu^2}\sum_{i=1}^{n}\{(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^{\top}Z_i\}^2u_{n,i}^2} = 2\sqrt{\mathrm{I}_n}\cdot\sqrt{\mathrm{II}_n} = o_{\mathcal{P}}(1)$$

by Lemmas 3.3 and 3.7. The result follows by the uniform version of Slutsky's theorem.

3.8.3 Proof of Proposition 3.3

As in the proof of Proposition 3.2, we work on the event that at least one $u_{n,1},\ldots,u_{n,n}$ is non-zero, which is a set of uniform asymptotic probability 1 by Assumption 3.2(b). Recall the definitions of ν from the proof of Proposition 3.2, and δ_{bias} from just after (3.9). Our test statistic can be written as

$$T_{\mathrm{Lasso}} = \frac{\frac{1}{\sqrt{n\nu}}\sum_{i=1}^{n}\varepsilon_iu_{n,i} - \frac{1}{\sqrt{n\nu}}\delta_{\mathrm{bias}}}{\sqrt{\frac{1}{ns_n^2}\sum_{i=1}^{n}(Y_i-\widehat{\boldsymbol{\gamma}}^{\top}Z_i)^2u_{n,i}^2 - \bigl(\frac{1}{n\nu}\sum_{i=1}^{n}\varepsilon_iu_{n,i} + \frac{1}{n\nu}\delta_{\mathrm{bias}}\bigr)^2}}.$$

Following the proof of Proposition 3.2, we know that $\frac{1}{\sqrt{n}\nu}\sum_{i=1}^n \varepsilon_i u_{n,i}$ converges uniformly to $N(0,1)$. Further, by Assumption 3.2(a) and Hölder's inequality, we have

$$\left|\frac{1}{\sqrt{n}\nu}\delta_{\text{bias}}\right| \leq \frac{1}{c^{1/2}}\left|\sum_{i=1}^n (\widehat{\gamma}-\gamma)^\top Z_i \upsilon_{n,i}\right| \leq \frac{1}{c^{1/2}}\|\widehat{\gamma}-\gamma\|_1 \left\|\sum_{i=1}^n Z_i \upsilon_{n,i}\right\|_\infty = o_{\mathcal{P}_0}(1)$$

under condition (3.10). A uniform version of Slutsky's theorem (Bengs and Holzmann, 2019, Theorem 6.3) now yields that the numerator of T_{Lasso} is uniformly asymptotically standard Gaussian. We can repeat the arguments of Proposition 3.2 to show that the denominator is $1+o_{\mathcal{P}_0}(1)$ under Assumption 3.2, so the uniform version of Slutsky's theorem yields the desired result.

3.8.4 Proof of Theorem 3.1

We prove the result under the given assumptions but instead of assuming that $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ are formed on an auxiliary sample, we let $R_{ij} := \mathbb{E}(M_i M_j \,|\, (X_i, Z_i)_{i=1}^n) - \mathbb{E}(M_i M_j \,|\, (Z_i)_{i=1}^n)$ and assume that

$$\frac{1}{n\sigma_n^2}\sum_{i\neq j}|\mathbb{E}(R_{ij}\xi_i\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f})| = o_{\mathcal{P}_0}(1). \tag{3.22}$$

In Proposition 3.6 we show that this condition is trivially satisfied if $\widehat{m}$ is formed out of sample. We also show that this is trivially satisfied if $\widehat{m}$ is a linear smoother or if $X \perp\!\!\!\perp Y \,|\, Z$. Define $\nu^2 := \text{Var}(\varepsilon\xi)$ and note that $\nu^2 \geq c\sigma^2$. Throughout this proof we work on the event $\Omega_0 := \{\sigma \neq 0\}$, which satisfies $\mathbb{P}(\Omega_0^c) = o_{\mathcal{P}_0}(1)$ by Assumption 3.3(a). Define

$$T^{(\text{N})} := \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n L_i}{\nu} \quad \text{and} \quad T^{(\text{D})} := \frac{\sqrt{\frac{1}{n}\sum_{i=1}^n L_i^2 - \left(\frac{1}{n}\sum_{i=1}^n L_i\right)^2}}{\nu},$$

so that $T = T^{(\text{N})}/T^{(\text{D})}$. We will show that $T^{(\text{N})}$ converges uniformly in distribution to $\mathcal{N}(0,1)$ and $|T^{(\text{D})}-1| = o_{\mathcal{P}_0}(1)$, which yields the desired result by combining Lemma 3.7 and the uniform version of Slutsky's lemma (Bengs and Holzmann, 2019, Theorem 6.3).

Define $M_i := m(Z_i) - \widehat{m}(Z_i)$ and $\widetilde{M}_i := m_{\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i)$ for $i \in [n]$ and note that

$$T^{(\text{N})} = \underbrace{\frac{1}{\sqrt{n}\nu}\sum_{i=1}^n M_i\widetilde{M}_i}_{a_n} + \underbrace{\frac{1}{\sqrt{n}\nu}\sum_{i=1}^n \widetilde{M}_i\varepsilon_i}_{b_n} + \underbrace{\frac{1}{\sqrt{n}\nu}\sum_{i=1}^n M_i\xi_i}_{c_n} + \underbrace{\frac{1}{\sqrt{n}\nu}\sum_{i=1}^n \varepsilon_i\xi_i}_{U_n}. \tag{3.23}$$

By the Cauchy–Schwarz inequality,

$$a_n \leq \sqrt{\frac{1}{cn}\left(\sum_{i=1}^n M_i^2\right)\left(\frac{1}{\sigma^2}\sum_{i=1}^n \widetilde{M}_i^2\right)} = o_{\mathcal{P}_0}(1),$$

by Assumption 3.3(b).

To see that $b_n = o_{\mathcal{P}_0}(1)$, we note that

$$b_n^2 = \frac{1}{n\nu^2}\sum_{i=1}^n \widetilde{M}_i^2\varepsilon_i^2 + \frac{1}{n\nu^2}\sum_{i\neq j}\widetilde{M}_i\widetilde{M}_j\varepsilon_i\varepsilon_j \tag{3.24}$$

and the first term is $o_{\mathcal{P}_0}(1)$ by Assumption 3.3(c). Now, for $i \neq j$,

$$\mathbb{E}(Y_iY_j \,|\, X_i, X_j, Z_i, Z_j) = \mathbb{E}(Y_i \,|\, X_i, Z_i)\mathbb{E}(Y_j \,|\, X_j, Z_j) = m(Z_i)m(Z_j),$$

using the fact that $m(Z) = \mathbb{E}(Y \,|\, Z) = \mathbb{E}(Y \,|\, X, Z)$ under $\mathcal{P}_0$. Hence,

$$\begin{aligned}\mathbb{E}(\varepsilon_i\varepsilon_j \,|\, X_i, X_j, Z_i, Z_j) &= \mathbb{E}\big\{\big(Y_i - m(Z_i)\big)\big(Y_j - m(Z_j)\big) \,|\, X_i, X_j, Z_i, Z_j\big\} \\ &= \mathbb{E}(Y_iY_j \,|\, X_i, X_j, Z_i, Z_j) - m(Z_i)m(Z_j) = 0.\end{aligned}$$

It follows that

$$\begin{aligned}&\mathbb{E}\Big(\frac{1}{n\nu^2}\sum_{i\neq j}\widetilde{M}_i\widetilde{M}_j\varepsilon_i\varepsilon_j \;\Big|\; (X_i, Z_i)_{i=1}^n, \widehat{f}, \widehat{m}_{\widehat{f}}\Big) \\ &\qquad = \frac{1}{n\nu^2}\sum_{i\neq j}\widetilde{M}_i\widetilde{M}_j\mathbb{E}(\varepsilon_i\varepsilon_j \,|\, X_i, X_j, Z_i, Z_j) = 0,\end{aligned}$$

and we deduce by Lemmas 3.6 and 3.7 that $b_n = o_{\mathcal{P}_0}(1)$.

To see that $c_n = o_{\mathcal{P}_0}(1)$, we proceed as above and write

$$c_n^2 = \frac{1}{n\nu^2}\sum_{i=1}^n M_i^2\xi_i^2 + \frac{1}{n\nu^2}\sum_{i\neq j}M_iM_j\xi_i\xi_j,$$

where we again note that the first term is $o_{\mathcal{P}_0}(1)$ by Assumption 3.3(c). Moreover,

$$\begin{aligned}&\mathbb{E}\Big(\frac{1}{n\nu^2}\sum_{i\neq j}M_iM_j\xi_i\xi_j \;\Big|\; (Z_i)_{i=1}^n, \widehat{f}\Big) \\ &\qquad = \frac{1}{n\nu^2}\sum_{i\neq j}\mathbb{E}\big\{\mathbb{E}\big(M_iM_j \,|\, (X_i, Z_i)_{i=1}^n\big)\xi_i\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f}\big\} \\ &\qquad = \frac{1}{n\nu^2}\sum_{i\neq j}\mathbb{E}(R_{ij}\xi_i\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f}),\end{aligned}$$

where the last equality holds since

$$\begin{aligned}\mathbb{E}\big\{\mathbb{E}\big(M_iM_j \,|\, (Z_i)_{i=1}^n\big)\xi_i\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f}\big\} &= \mathbb{E}\big(M_iM_j \,|\, (Z_i)_{i=1}^n\big)\mathbb{E}(\xi_i \,|\, Z_i, \widehat{f})\mathbb{E}(\xi_j \,|\, Z_j, \widehat{f}) \\ &= 0.\end{aligned}$$

Continuing, we have by (3.22) that

$$\frac{1}{n\nu^2}\sum_{i\neq j}\mathbb{E}(R_{ij}\xi_i\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f}) \leq \frac{1}{cn\sigma^2}\sum_{i\neq j}|\mathbb{E}(R_{ij}\xi_{n,i}\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f})| = o_{\mathcal{P}_0}(1).$$

Therefore, by Lemmas 3.6 and 3.7 we conclude that $c_n = o_{\mathcal{P}_0}(1)$ as desired.

To deal with the final term, we define the triangular array $W_{n,i} := \nu^{-1}\varepsilon_i\xi_i$ for $n \in \mathbb{N}$ and $i \in [n]$, and note that $W_{n,i}$ satisfies all the conditions of Lemma 3.8 by Assumptions 3.3(a) and (d) (here we condition on $\widehat{f}$ in applying this result). Hence, $U_n = n^{-1/2}\sum_{i=1} W_{n,i}$, and therefore $T^{(\mathrm{N})}$, converges uniformly in distribution to $\mathcal{N}(0,1)$.

We now show that $|(T^{(\mathrm{D})})^2 - 1| = o_{\mathcal{P}_0}(1)$, from which the desired result follows from Lemma 3.7. Note that

$$(T^{(\mathrm{D})})^2 = \underbrace{\frac{1}{n\nu^2}\sum_{i=1}^n L_i^2}_{p_n} - \underbrace{\Big(\frac{1}{\sqrt{n}\nu}\sum_{i=1}^n L_i\Big)^2}_{q_n} \tag{3.25}$$

and that $q_n = \frac{1}{\sqrt{n}}T^{(\mathrm{N})}$. We have just shown that $T^{(\mathrm{N})} = O_{\mathcal{P}_0}(1)$, so $q_n = o_{\mathcal{P}_0}(1)$ and we are therefore done if we can show that $|p_n - 1| = o_{\mathcal{P}_0}(1)$. Now

$$\begin{aligned} p_n = {} & \underbrace{\frac{1}{n\nu^2}\sum_{i=1}^n \varepsilon_i^2\xi_i^2}_{\mathrm{I}_n} + \underbrace{\frac{1}{n\nu^2}\sum_{i=1}^n M_i^2\widetilde{M}_i^2}_{\mathrm{II}_n} + \underbrace{\frac{4}{n\nu^2}\sum_{i=1}^n M_i\widetilde{M}_i\varepsilon_i\xi_i}_{\mathrm{III}_n} + \underbrace{\frac{1}{n\nu^2}\sum_{i=1}^n \widetilde{M}_i^2\varepsilon_i^2}_{\mathrm{IV}_n^\varepsilon} \\ & + \underbrace{\frac{1}{n\nu^2}\sum_{i=1}^n M_i^2\xi_i^2}_{\mathrm{IV}_n^\xi} + \underbrace{\frac{2}{n\nu^2}\sum_{i=1}^n \widetilde{M}_i^2 M_i\varepsilon_i}_{\mathrm{V}_n^\varepsilon} + \underbrace{\frac{2}{n\nu^2}\sum_{i=1}^n M_i^2\widetilde{M}_i\xi_i}_{\mathrm{V}_n^\xi} \\ & + \underbrace{\frac{2}{n\nu^2}\sum_{i=1}^n \widetilde{M}_i\xi_i\varepsilon_i^2}_{\mathrm{VI}_n^\varepsilon} + \underbrace{\frac{2}{n\nu^2}\sum_{i=1}^n M_i\varepsilon_i\xi_i^2}_{\mathrm{VI}_n^\xi}. \end{aligned} \tag{3.26}$$

Consider the triangular array $\tilde{W}_{n,i} := W_{n,i}^2$ for $n \in \mathbb{N}$ and $i \in [n]$, and note that it satisfies all the conditions of Lemma 3.9 by Assumptions 3.3(a) and (d) with $\mu_n = 1$ (again conditioning on $\widehat{f}$ in that result), so $|\mathrm{I}_n - 1| = o_{\mathcal{P}_0}(1)$. It remains to show that the remaining terms are $o_{\mathcal{P}_0}(1)$. Now

$$0 \leq \mathrm{II}_n \leq \frac{1}{cn}\Big(\sum_{i=1}^n M_i^2\Big)\Big(\frac{1}{\sigma^2}\sum_{i=1}^n \widetilde{M}_i^2\Big) = o_{\mathcal{P}_0}(1)$$

by Assumption 3.3(b). By the Cauchy–Schwarz inequality,

$$|\mathrm{III}_n| \leq 4\Big(\frac{1}{n\nu^2}\sum_{i=1}^n \varepsilon_i^2\xi_i^2\Big)^{1/2}\Big(\frac{1}{n\nu^2}\sum_{i=1}^n M_i^2\widetilde{M}_i^2\Big)^{1/2} = 4\mathrm{I}_n^{1/2}\mathrm{II}_n^{1/2} = o_{\mathcal{P}_0}(1)$$

by the above and Lemma 3.7. Now

$$|\mathrm{IV}_n^\varepsilon| \le \frac{1}{n\nu^2}\sum_{i=1}^n \widetilde{M}_i^2\varepsilon_i^2 = o_{\mathcal{P}_0}(1)$$

by Assumption 3.3(c). A similar argument shows that $\mathrm{IV}_n^\xi = o_{\mathcal{P}_0}(1)$. By the triangle inequality and the Cauchy–Schwarz inequality,

$$|\mathrm{V}_n^\varepsilon| \le 2\Big(\frac{1}{n\nu^2}\sum_{i=1}^n M_i^2\widetilde{M}_i^2\Big)^{1/2}\Big(\frac{1}{n\nu^2}\sum_{i=1}^n \widetilde{M}_i^2\varepsilon_i^2\Big)^{1/2} = 2\mathrm{II}_n^{1/2}(\mathrm{IV}_n^\varepsilon)^{1/2} = o_{\mathcal{P}_0}(1)$$

by the above and Lemma 3.7. A similar argument can be used for V_n^ξ. Finally, again by the triangle inequality and the Cauchy–Schwarz inequality,

$$|\mathrm{VI}_n^\varepsilon| \le 2\Big(\frac{1}{n\nu^2}\sum_{i=1}^n \varepsilon_i^2\xi_i^2\Big)^{1/2}\Big(\frac{1}{n\nu^2}\sum_{i=1}^n \widetilde{M}_i^2\varepsilon_i^2\Big)^{1/2} = 2\mathrm{I}_n^{1/2}(\mathrm{IV}_n^\varepsilon)^{1/2} = o_{\mathcal{P}_0}(1)$$

by the above and Lemma 3.7; VI_n^ξ is handled similarly.

3.8.5 Proof of Theorem 3.2

Without loss of generality, we may assume that $\alpha \in (0, 1/2)$, so that $z_{1-\alpha} > 0$. Let s denote the denominator in the definition of T. Suppose there exists $c > 0$ such that

$$\sup_{P\in\mathcal{P}_1(\epsilon_n)} \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n L_i \le c\tau\Big) \to 0, \tag{3.27}$$

$$\frac{s}{\sqrt{n}} = o_{\mathcal{P}_1(\epsilon_n)}(\tau). \tag{3.28}$$

Note that, since $0/0 := 0$ and $\tau > 0$, we have that

$$\begin{aligned}\mathbb{P}(T \le z_{1-\alpha}) &= \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n L_i \le z_{1-\alpha}\frac{s}{\sqrt{n}}\Big)\\ &\le \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n L_i \le c\tau\Big) + \mathbb{P}\Big(z_{1-\alpha}\frac{s}{\sqrt{n}\tau} \ge c\Big).\end{aligned}$$

Thus, from (3.27) and (3.28),

$$\sup_{P\in\mathcal{P}_1(\epsilon_n)} \mathbb{P}(T \le z_{1-\alpha}) \le \sup_{P\in\mathcal{P}_1(\epsilon_n)} \left\{\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n L_i \le c\tau\Big) + \mathbb{P}\Big(z_{1-\alpha}\frac{s_n}{\sqrt{n}\tau} \ge c\Big)\right\} \to 0$$

and hence the result will follow if we can prove (3.27) and (3.28).

Observe that if we define

$$\check{f}(X, Z) := \frac{\tau^{1/2}}{\sigma^{1/2}}\widehat{f}(X, Z) \tag{3.29}$$

and let $\check{T}$ denote the test using $\check{f}$ in place of $\widehat{f}$, then $T = \check{T}_n$, since we have assumed that $m_{\widehat{f}}$ is scale equivariant. It follows that we may put $\check{f}$ in place of $\widehat{f}$ and assume without loss of generality that

$$\sigma = \mathbb{E}(\xi^2 \,|\, \widehat{f}) = \tau.$$

For both claims (3.27) and (3.28), we therefore work with $\check{f}$ instead of $\widehat{f}$.

To prove (3.27), we write $Y_i = m(Z_i) + h(X_i, Z_i) + \zeta_i$, and have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n L_i = &\underbrace{\frac{1}{n}\sum_{i=1}^n h(X_i, Z_i)\xi_i}_{\mathrm{I}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^n \zeta_i\xi_i}_{\mathrm{II}_n} \\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^n \{m(Z_i) - \widehat{m}(Z_i)\}\xi_i}_{\mathrm{III}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^n h(X_i, Z_i)\{m_{\check{f}}(Z_i) - \widehat{m}_{\check{f}}(Z_i)\}}_{\mathrm{IV}_n} \\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^n \zeta_i\{m_{\check{f}}(Z_i) - \widehat{m}_{\check{f}}(Z_i)\}}_{\mathrm{V}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^n \{m(Z_i) - \widehat{m}(Z_i)\}\{m_{\check{f}}(Z_i) - \widehat{m}_{\check{f}}(Z_i)\}}_{\mathrm{VI}_n}.\end{aligned}$$

Defining the triangular array $W_{n,i} := h(X_i, Z_i)\xi_i/\tau$ for $i \in [n]$, note that

$$\sum_{i=1}^n \mathbb{E}(|W_{n,i}|^2 \,|\, \check{f}) = \frac{n\mathbb{E}(h(X,Z)^2\xi^2 \,|\, \check{f})}{\tau^2} \leq \frac{C^2 n}{\tau}.$$

Therefore, defining $\mu_n := \mathbb{E}(\mathrm{I}_n \,|\, \check{f})$ (the numerator of $\mathrm{Corr}(h(X,Z), \xi \,|\, \check{f})$), assumption *(ii)* of Lemma 3.9 is satisfied with $\delta = 1$ on $\mathcal{P}_1(\epsilon_n)$ by (3.15). We deduce that

$$\sup_{P \in \mathcal{P}_1(\epsilon_n)} \mathbb{P}(|\mathrm{I}_n - \mu_n| \geq \eta\tau) = o(1)$$

for any $\eta > 0$.

To deal with the II_n term, we first note that for $i \neq j$,

$$\mathbb{E}(\zeta_i\zeta_j \,|\, X_i, Z_i, X_j, Z_j) = \mathbb{E}(\zeta_i \,|\, X_i, Z_i)\mathbb{E}(\zeta_j \,|\, X_j, Z_j) = 0.$$

Hence, using the fact that $\mathbb{E}(\zeta_i^2 \,|\, X_i, Z_i) = \mathrm{Var}(Y_i \,|\, X_i, Z_i) \leq C$ for all $i \in [n]$, we have

$$\begin{aligned}\mathbb{E}(|\mathrm{II}_n| \,|\, \check{f}, (X_i, Z_i)_{i=1}^n) &\leq \frac{1}{n}\Big(\sum_{i=1}^n \mathbb{E}(\zeta_i^2 \,|\, X_i, Z_i)\xi_i^2\Big)^{1/2} \\ &\leq \frac{C^{1/2}}{n^{1/2}}\Big(\frac{1}{n}\sum_{i=1}^n \xi_i^2\Big)^{1/2},\end{aligned}$$

and therefore

$$\mathbb{E}(|\mathrm{II}_n| \,|\, \check{f}) \leq \frac{C^{1/2}}{n^{1/2}}\tau^{1/2}.$$

We conclude by Lemma 3.2 that $\mathrm{II}_n = O_{\mathcal{P}_1(\epsilon_n)}(n^{-1/2}\tau^{1/2})$. To deal with the III_n term, we note similarly that

$$\mathbb{E}(\xi_i\xi_j \mid Z_i, Z_j) = \mathbb{E}(\xi_i \mid Z_i)\mathbb{E}(\xi_j \mid Z_j) = 0. \tag{3.30}$$

Thus,

$$\begin{aligned}\mathbb{E}(|\mathrm{III}_n| \mid \check{f}, (Z_i)_{i=1}^n, \widehat{m}) &\leq \frac{1}{n}\biggl(\sum_{i=1}^n \mathbb{E}(\xi_i^2 \mid Z_i)\{m(Z_i) - \widehat{m}(Z_i)\}^2\biggr)^{1/2} \\ &= \frac{\tau^{1/2}}{n^{1/2}}\biggl(\frac{1}{n}\sum_{i=1}^n \{m(Z_i) - \widehat{m}(Z_i)\}^2\biggr)^{1/2}.\end{aligned}$$

We deduce by Lemma 3.2 and Assumption 3.4(a) that $\mathrm{III}_n = O_{\mathcal{P}_1(\epsilon_n)}(\tau^{1/2}n^{-(\beta_1+1)/2})$. Note further that

$$\frac{1}{n}\sum_{i=1}^n \{m_{\check{f}}(Z_i) - \widehat{m}_{\check{f}}(Z_i)\}^2 = \frac{\tau}{n\sigma^2}\sum_{i=1}^n \{m_{\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i)\}^2 = O_{\mathcal{P}_1(\epsilon_n)}(\tau n^{-\beta_2}), \tag{3.31}$$

by Assumption 3.4(b). We can thus repeat the calculation in (3.30) for h, letting us show that

$$\begin{aligned}\mathbb{E}(|\mathrm{IV}_n| \mid \check{f}, (Z_i)_{i=1}^n, \widehat{m}_{\check{f}}) &\leq \frac{1}{n}\biggl(\sum_{i=1}^n \mathbb{E}(h(X_i, Z_i)^2 \mid Z_i)\{m_{\check{f}}(Z_i) - \widehat{m}_{\check{f}}(Z_i)\}^2\biggr)^{1/2} \\ &= \frac{\tau^{1/2}}{n^{1/2}}\biggl(\frac{1}{n}\sum_{i=1}^n \{m_{\check{f}}(Z_i) - \widehat{m}_{\check{f}}(Z_i)\}^2\biggr)^{1/2}.\end{aligned}$$

Thus, Lemma 3.2 and (3.31) let us conclude that $\mathrm{IV}_n = O_{\mathcal{P}_1(\epsilon_n)}(\tau n^{-(\beta_2+1)/2})$. For the V_n term, we note that by similar arguments as above,

$$\mathbb{E}(|\mathrm{V}_n| \mid \check{f}, (X_i, Z_i)_{i=1}^n) \leq \frac{C^{1/2}}{n^{1/2}}\biggl(\frac{1}{n}\sum_{i=1}^n \{m_{\check{f}}(Z_i) - \widehat{m}_{\check{f}}(Z_i)\}^2\biggr)^{1/2}. \tag{3.32}$$

Hence, by Lemma 3.2 and (3.31), we deduce that $\mathrm{V}_n = O_{\mathcal{P}_1(\epsilon_n)}(\tau^{1/2}n^{-(\beta_2+1)/2})$. For the final term, by the Cauchy–Schwarz inequality and (3.31),

$$\begin{aligned}|\mathrm{VI}_n| &\leq \biggl(\frac{1}{n}\sum_{i=1}^n \{m(Z_i) - \widehat{m}(Z_i)\}^2\biggr)^{1/2}\biggl(\frac{1}{n}\sum_{i=1}^n \{m_{\check{f}}(Z_i) - \widehat{m}_{\check{f}}(Z_i)\}^2\biggr)^{1/2} \\ &= O_{\mathcal{P}_1(\epsilon_n)}(\tau^{1/2}n^{-(\beta_1+\beta_2)/2})\end{aligned}$$

using Assumptions 3.4(a) and (3.31). Letting $R_n := \mathrm{II}_n + \mathrm{III}_n + \mathrm{IV}_n + \mathrm{V}_n + \mathrm{VI}_n$, we have therefore shown that $R_n = o_{\mathcal{P}_1(\epsilon_n)}(\tau)$ by (3.15). We conclude that

$$\sup_{P\in\mathcal{P}_1(\epsilon_n)} \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n L_{n,i} \le \rho\tau/3\Big) \le \sup_{P\in\mathcal{P}_1(\epsilon_n)} \mathbb{P}(\mu_n \le \rho\tau) + \sup_{P\in\mathcal{P}_1(\epsilon_n)} \mathbb{P}(|\mathrm{I}_n - \mu_n| \ge \rho\tau/3) + \sup_{P\in\mathcal{P}_1(\epsilon_n)} \mathbb{P}(|R_n| \ge \rho\tau/3),$$

so (3.27) is satisfied with $c := \rho/3$ by Assumption 3.4(c).

To see that (3.28) holds, note that

$$\begin{aligned}\frac{s_n}{n^{1/2}} \le \Big(\frac{1}{n^2}\sum_{i=1}^n L_i^2\Big)^{1/2} \le 5^{1/2}\Bigg[&\underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n h(X_i,Z_i)^2\xi_i^2\Big)^{1/2}}_{\widetilde{\mathrm{I}}_n} \\ &+ \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \zeta_i^2\xi_i^2\Big)^{1/2}}_{\widetilde{\mathrm{II}}_n} + \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \{m(Z_i)-\widehat{m}(Z_i)\}^2\xi_i^2\Big)^{1/2}}_{\widetilde{\mathrm{III}}_n} \\ &+ \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n h(X_i,Z_i)^2\{m_{\check{f}}(Z_i)-\widehat{m}_{\check{f}}(Z_i)\}^2\Big)^{1/2}}_{\widetilde{\mathrm{IV}}_n} \\ &+ \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \zeta_i^2\{m_{\check{f}}(Z_i)-\widehat{m}_{\check{f}}(Z_i)\}^2\Big)^{1/2}}_{\widetilde{\mathrm{V}}_n} \\ &+ \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \{m(Z_i)-\widehat{m}(Z_i)\}^2\{m_{\check{f}}(Z_i)-\widehat{m}_{\check{f}}(Z_i)\}^2\Big)^{1/2}}_{\widetilde{\mathrm{VI}}_n}\Bigg].\end{aligned}$$

Now

$$\mathbb{E}(\widetilde{\mathrm{I}}_n \mid \check{f}) \le \frac{C}{n^{1/2}}\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\xi_i^2 \mid \check{f})\Big)^{1/2} = \frac{C}{n^{1/2}}\tau^{1/2},$$

so by Lemma 3.2 we see that $\widetilde{\mathrm{I}}_n = O_{\mathcal{P}_1(\epsilon_n)}(n^{-1/2}\tau^{1/2})$. The remaining terms are of the same uniform stochastic order as the corresponding terms without tildes using the bounds above. Thus, (3.28) is satisfied by (3.15), and the result follows.

If $\widehat{m}$ and $\widehat{m}_{\widehat{f}}$ are formed on $\mathcal{D}_1$, then only the III_n and IV_n terms will change. By the Cauchy–Schwarz inequality, Lemma 3.2 and (3.31), these terms satisfy

$$|\mathrm{III}_n| \le \Big(\frac{1}{n}\sum_{i=1}^n \{m(Z_i)-\widehat{m}(Z_i)\}^2\Big)^{1/2}\Big(\frac{1}{n}\sum_{i=1}^n \xi_i^2\Big)^{1/2} = O_{\mathcal{P}_1(\epsilon_n)}(n^{-\beta_1/2}\tau^{1/2})$$

and

$$|\mathrm{IV}_n| \le \left(\frac{1}{n}\sum_{i=1}^n h(X_i,Z_i)^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n\{m_{\check f}(Z_i)-\widehat m_{\check f}(Z_i)\}^2\right)^{1/2} = O_{\mathcal{P}_1(\epsilon_n)}(n^{-\beta_2/2}\tau^{1/2})$$

thus we would need $\epsilon_n \cdot n^{\min\{1,\beta_1,\beta_2\}} \to \infty$ rather than (3.15) to prove the result.

3.8.6 Proof of Theorem 3.3

It suffices to check the conditions of Assumption 3.3 as the result will then follow by Theorem 3.1.

By Proposition 3.13 with $\widehat f$ in place of f in that result,

$$\xi := \widehat f(X,Z) - \mathbb{E}\{\widehat f(X,Z)\,|\,Z,\widehat f\} = (\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}})^\top\{\boldsymbol{\phi}(X,Z) - \mathbb{E}(\boldsymbol{\phi}(X,Z)\,|\,Z)\}. \tag{3.33}$$

Thus, from the definition in (3.12),

$$\sigma^2 = \mathbb{E}(\xi^2\,|\,\widehat f) = (\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}})^\top\boldsymbol{\Lambda}(\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}) \ge \tilde\lambda_{\min}(\boldsymbol{\Lambda})\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_2^2 \ge cK_{XZ}^{-1}\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_2^2, \tag{3.34}$$

where the last inequality holds by our assumption. Hence,

$$\sup_{P\in\mathcal{P}_0}\mathbb{P}(\sigma^2=0) = \sup_{P\in\mathcal{P}_0}\mathbb{P}_P(\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty = 0) = o(1),$$

so Assumption 3.3(a) is satisfied.

Next, by Corollary 3.2,

$$\frac{1}{n}\sum_{i=1}^n M_i^2 = O_{\mathcal{P}_0}(\widetilde K_Z^{-2s/d_Z} + \widetilde K_Z/n). \tag{3.35}$$

Now, suppose that $g^\dagger = \boldsymbol{\beta}_{XZ}^\top\boldsymbol{\phi}$ is the $L_2(P)$-best approximant of g over $\mathcal{S}_{r,L}^d$. Then, by Propositions 3.13 and 3.9(b),

$$\begin{aligned}\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty &= \|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}_{XZ}\|_\infty \le 2\|\widehat{\boldsymbol{\beta}}_{XZ}\|_\infty \\ &\le 2\|\widehat{\boldsymbol{\beta}}_{XZ}-\boldsymbol{\beta}_{XZ}\|_\infty + 2c_s(r)^{-d}\|g-g^\dagger\|_\infty + 2c_s(r)^{-d}\|g\|_\infty.\end{aligned} \tag{3.36}$$

Hence, by Corollary 3.2, Propositions 3.10 and 3.11 and Assumption 3.5,

$$\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty = O_{\mathcal{P}_0}(K_{XZ}n^{-1/2}+1) = O_{\mathcal{P}_0}(1), \tag{3.37}$$

where the last equality uses (3.18). Now $\widetilde m(Z)$ is in the span of $\boldsymbol{\psi}(Z)$, so the residuals $m_{\widehat f}-\widehat m_{\widehat f}$ are identical to those resulting from a $\widehat g(X,Z)$ on $\boldsymbol{\psi}(Z)$ regression. Thus, by Proposition 3.14,

$$\frac{1}{n}\sum_{i=1}^n \widetilde M_i^2 = O_{\mathcal{P}_0}(\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty^2\{\widetilde K_Z^{-2s/d_Z} + \widetilde K_Z/n\}). \tag{3.38}$$

Combining (3.34), (3.35) and (3.38), we have

$$\left\{\frac{1}{n}\sum_{i=1}^n M_i^2\right\}\left\{\frac{1}{n\sigma^2}\sum_{i=1}^n \widetilde{M}_i^2\right\} = O_{\mathcal{P}_0}\left(\frac{\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty^2}{\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_2^2} K_{XZ}\{\widetilde{K}_Z^{-2s/d_Z} + \widetilde{K}_Z/n\}^2\right) = o_{\mathcal{P}_0}(n^{-1}),$$

by (3.17), so Assumption 3.3(b) holds.

For any $\eta \geq 1$, we have by (3.33) that

$$\mathbb{E}(|\xi|^\eta \mid Z, \widehat{f}) \leq 2^\eta \mathbb{E}(|(\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}})^\top \boldsymbol{\phi}(X,Z)|^\eta \mid Z, \widehat{f}) \leq 2^\eta \|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty^\eta \tag{3.39}$$

by Hölder's inequality and Proposition 3.9(a). Hence, taking $\eta = 2$, the first part of Assumption 3.3(c) is satisfied by Lemma 3.2, because

$$\begin{aligned}\mathbb{E}\left(\frac{1}{\sigma^2}\frac{1}{n}\sum_{i=1}^n M_i^2\xi_i^2 \;\middle|\; \widehat{f}, (Z_i)_{i=1}^n, \widehat{m}\right) &= \frac{1}{\sigma^2}\left\{\frac{1}{n}\sum_{i=1}^n M_i^2 \mathbb{E}(\xi_i^2 \mid Z_i, \widehat{f})\right\} \\ &= O_{\mathcal{P}_0}\left(\frac{\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty^2}{\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_2^2} K_{XZ}\{\widetilde{K}_Z^{-2s/d_Z} + \widetilde{K}_Z/n\}\right) = o_{\mathcal{P}_0}(1).\end{aligned}$$

Moreover, since $\mathbb{E}(\varepsilon^2 \mid X, Z) \leq \mathbb{E}(\varepsilon^{2+\delta} \mid X, Z)^{2/(2+\delta)} \leq C^{2/(2+\delta)}$, we have

$$\begin{aligned}\mathbb{E}\left(\frac{1}{\sigma^2}\frac{1}{n}\sum_{i=1}^n \widetilde{M}_i^2\varepsilon_i^2 \;\middle|\; \widehat{f}, (X_i, Z_i)_{i=1}^n, \widehat{m}_{\widehat{f}}\right) &= \frac{1}{\sigma^2}\left\{\frac{1}{n}\sum_{i=1}^n \widetilde{M}_i^2 \mathbb{E}(\varepsilon_i^2 \mid X_i, Z_i)\right\} \\ = O_{\mathcal{P}_0}\left(\frac{\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty^2}{\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_2^2} K_{XZ}\{\widetilde{K}_Z^{-2s/d_Z} + \widetilde{K}_Z/n\}\right) &= o_{\mathcal{P}_0}(K_{XZ}^{1/2} n^{-1/2}) = o_{\mathcal{P}_0}(1)\end{aligned}$$

by (3.17) and (3.18), so the second part of Assumption 3.3(c) holds by Lemma 3.2. Finally, by (3.34), Assumption 3.5(a), (3.39) with $\eta = 2 + \delta$ and (3.18), we have

$$\frac{\mathbb{E}(|\xi\varepsilon|^{2+\delta} \mid \widehat{f})}{\sigma^{2+\delta}} \leq \frac{2^{2+\delta}C}{c^{2+\delta}} \frac{\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty^{2+\delta}}{\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_2^{2+\delta}} K_{XZ}^{1+\delta/2} \leq \frac{2^{2+\delta}C}{c^{2+\delta}} K_{XZ}^{1+\delta/2} = o(n^{\delta/2}),$$

so Assumption 3.3(d) is satisfied. This establishes the claim.

3.8.7 Proof of Theorem 3.4

Without loss of generality, we may assume that $\alpha \in (0, 1/2)$, so that $z_{1-\alpha} > 0$. Let s_n denote the denominator in the definition of T_{spline}. Suppose we can show that

$$\frac{1}{n}\sum_{i=1}^n L_i = \tau(1 + R_n), \quad \text{where } R_n = o_{\mathcal{P}_1(\epsilon_n)}(1) \tag{3.40}$$

$$\frac{s}{\sqrt{n}} = \tau U_n, \quad \text{where } U_n = o_{\mathcal{P}_1(\epsilon_n)}(1). \tag{3.41}$$

Note that, since $0/0 := 0$ and $\tau > 0$, we have from (3.40) and (3.41) that

$$\begin{aligned}\mathbb{P}(T_{\text{Spline}} \le z_{1-\alpha}) &= \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n L_i \le z_{1-\alpha}\frac{s}{\sqrt{n}}\Big) = \mathbb{P}(z_{1-\alpha}U_n - R_n \ge 1)\\ &\le \mathbb{P}\Big(|U_n| \ge \frac{1}{2z_{1-\alpha}}\Big) + \mathbb{P}\Big(|R_n| \ge \frac{1}{2}\Big).\end{aligned}$$

Thus, from (3.40) and (3.41),

$$\sup_{P\in\mathcal{P}_1(\epsilon_n)} \mathbb{P}(T_{\text{Spline}} \le z_{1-\alpha}) \le \sup_{P\in\mathcal{P}_1(\epsilon_n)} \mathbb{P}\Big(|U_n| \ge \frac{1}{2z_{1-\alpha}}\Big) + \sup_{P\in\mathcal{P}_1(\epsilon_n)} \mathbb{P}\Big(|R_n| \ge \frac{1}{2}\Big) \to 0$$

and hence the result will follow if we can prove (3.40) and (3.41).

To see that (3.40) holds, we write

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n L_i &= \underbrace{\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i,Z_i)}_{\mathrm{I}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^n \varepsilon_i\{m(Z_i)-\widetilde{m}(Z_i)\}}_{\mathrm{II}_n}\\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^n f(X_i,Z_i)\{\widehat{g}(X_i,Z_i)-g(X_i,Z_i)\}}_{\mathrm{III}_n}\\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^n \{Y_i-g(X_i,Z_i)\}\{\widehat{g}(X_i,Z_i)-g(X_i,Z_i)\}}_{\mathrm{IV}_n} - \underbrace{\frac{1}{n}\sum_{i=1}^n \varepsilon_i m_{\widehat{f}}(Z_i)}_{\mathrm{V}_n}\\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^n \varepsilon_i\{m_{\widehat{f}}(Z_i)-\widehat{m}_{\widehat{f}}(Z_i)\}}_{\mathrm{VI}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^n \{m(Z_i)-\widehat{m}(Z_i)\}\{m_{\widehat{f}}(Z_i)-\widehat{m}_{\widehat{f}}(Z_i)\}}_{\mathrm{VII}_n}\\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^n \{m(Z_i)-\widehat{m}(Z_i)\}\{\widehat{f}(X_i,Z_i)-m_{\widehat{f}}(Z_i)\}}_{\mathrm{VIII}_n}.\end{aligned}$$

Using the fact that $(\varepsilon_i f(X_i,Z_i))_{i=1}^n$ are independent and identically distributed with mean τ, we have that

$$\begin{aligned}&\mathbb{E}\Big(\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i,Z_i) - \tau\Big|\Big) \le \mathbb{E}\Big\{\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i,Z_i) - \tau\Big)^2\Big\}^{1/2}\\ &= \frac{1}{n^{1/2}}\{\mathrm{Var}(\varepsilon f(X,Z))\}^{1/2} \le \frac{1}{n^{1/2}}\{\mathbb{E}(\varepsilon^2 f(X,Z)^2)\}^{1/2} \le \Big(\frac{C^{2/(2+\delta)}\tau}{n}\Big)^{1/2},\end{aligned} \tag{3.42}$$

so $\mathrm{I}_n - \tau = O_{\mathcal{P}}(\tau^{1/2}/n^{1/2})$, by Lemma 3.2. Now note that for $i \ne j$,

$$\mathbb{E}(\varepsilon_i\varepsilon_j \,|\, Z_i, Z_j) = \mathbb{E}(\varepsilon_i \,|\, Z_i)\mathbb{E}(\varepsilon_j \,|\, Z_j) = 0. \tag{3.43}$$

Hence, by Assumption 3.5(a), we have that

$$\begin{aligned}\mathbb{E}(|\mathrm{II}_n| \,|\, \widetilde{m}, (Z_i)_{i=1}^n) &\leq \frac{1}{n}\Big(\sum_{i=1}^n \mathbb{E}(\varepsilon_i^2 \,|\, Z_i)\{m(Z_i) - \widetilde{m}(Z_i)\}^2\Big)^{1/2} \\ &\leq \frac{C^{1/(2+\delta)}}{n^{1/2}}\Big(\frac{1}{n}\sum_{i=1}^n \{m(Z_i) - \widetilde{m}(Z_i)\}^2\Big)^{1/2}.\end{aligned}$$

Thus, by Corollary 3.2 and Lemma 3.2,

$$\mathrm{II}_n = O_{\mathcal{P}}(K_Z^{-s/d_Z} n^{-1/2} + K_Z^{1/2} n^{-1}) = O_{\mathcal{P}}(n^{-\frac{(4s+d/2)}{4s+d}} + n^{-\frac{(4s+d_X)}{4s+d}}) = O_{\mathcal{P}}(n^{-\frac{4s}{4s+d}}).$$

Since $\tau = \mathbb{E}(f(X,Z)^2)$, we have by Proposition 3.16 that

$$\begin{aligned}\mathrm{III}_n &= O_{\mathcal{P}}(K_{XZ}^{-2s/d} + K_{XZ}^{-(s/d-1/2)} n^{-1} + \tau^{1/2} n^{-1/2}\{1 + K_{XZ}^{-(s/d-1/2)}\}) \\ &= O_{\mathcal{P}}(n^{-\frac{4s}{4s+d}} + \tau^{1/2} n^{-1/2}).\end{aligned}$$

Next, by Assumption 3.5(a) and (c),

$$\begin{aligned}\mathbb{E}(\{Y - g(X,Z)\}^2 \,|\, X, Z) &= \mathbb{E}(\{Y - m(Z)\}^2 \,|\, X, Z) - 2m(Z)^2 + f(X,Z)^2 \\ &\leq C^{2/(2+\delta)} + 4C^2\end{aligned}$$

and for $i \neq j$,

$$\begin{aligned}&\mathbb{E}(\{Y_i - g(X_i, Z_i)\}\{Y_j - g(X_j, Z_j)\} \,|\, X_i, Z_i, X_j, Z_j) \\ &= \mathbb{E}(\{Y_i - g(X_i, Z_i)\} \,|\, X_i, Z_i)\mathbb{E}(\{Y_j - g(X_j, Z_j)\} \,|\, X_j, Z_j) = 0.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}(|\mathrm{IV}_n| \,|\, \widehat{g}, (X_i, Z_i)_{i=1}^n) \\ &\leq \frac{1}{n}\Big(\sum_{i=1}^n \mathbb{E}(\{Y_i - g(X_i, Z_i)\}^2 \,|\, X_i, Z_i)\{\widehat{g}(X_i, Z_i) - g(X_i, Z_i)\}^2\Big)^{1/2} \\ &\leq \frac{(C^{2/(2+\delta)} + 4C^2)^{1/2}}{n^{1/2}}\Big(\frac{1}{n}\sum_{i=1}^n \{\widehat{g}(X_i, Z_i) - g(X_i, Z_i)\}^2\Big)^{1/2}.\end{aligned}$$

By Corollary 3.2 and Lemma 3.2 we thus have

$$\mathrm{IV}_n = O_{\mathcal{P}}(K_{XZ}^{-s/d} n^{-1/2} + K_{XZ}^{1/2} n^{-1}) = O_{\mathcal{P}}(n^{-\frac{(4s+d/2)}{4s+d}} + n^{-\frac{4s}{4s+d}}) = O_{\mathcal{P}}(n^{-\frac{4s}{4s+d}}).$$

Now, using (3.43) and the fact that $m_{\widehat{f}}(Z) = \mathbb{E}(\widehat{f}(X,Z) - f(X,Z) \,|\, Z, \widehat{f})$, we have

$$\begin{aligned}\mathbb{E}(|\mathrm{V}_n| \,|\, \widehat{f}, (Z_i)_{i=1}^n) &\leq \frac{C^{1/(2+\delta)}}{n^{1/2}} \bigg(\frac{1}{n} \sum_{i=1}^n \mathbb{E}(\{\widehat{f}(X_i,Z_i) - f(X_i,Z_i)\}^2 \,|\, Z_i, \widehat{f}) \bigg)^{1/2} \\ &\leq \frac{2^{1/2} C^{1/(2+\delta)}}{n^{1/2}} \bigg(\frac{1}{n} \sum_{i=1}^n \mathbb{E}(\{\widehat{g}(X_i,Z_i) - g(X_i,Z_i)\}^2 \,|\, Z_i, \widehat{f}) \\ &\qquad + \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\{\widetilde{m}(Z_i) - m(Z_i)\}^2 \,|\, Z_i, \widehat{f}) \bigg)^{1/2}.\end{aligned}$$

By Corollary 3.2 and Lemma 3.2, we deduce that

$$\mathrm{V}_n = O_{\mathcal{P}}(K_{XZ}^{-s/d} n^{-1/2} + K_{XZ}^{1/2} n^{-1} + K_Z^{-s/d_Z} n^{-1/2} + K_Z^{1/2} n^{-1}) = O_{\mathcal{P}}(n^{-\frac{4s}{4s+d}}).$$

As in the proof of Theorem 3.3, $\widetilde{m}(Z)$ is in the span of $\boldsymbol{\psi}(Z)$, the residuals $m_{\widehat{f}} - \widehat{m}_{\widehat{f}}$ are identical to those resulting from a $\widehat{g}(X,Z)$ on $\boldsymbol{\psi}(Z)$ regression. Moreover, by (3.37),

$$\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty = O_{\mathcal{P}}(K_{XZ} n^{-1/2} + 1) = O_{\mathcal{P}}(1), \tag{3.44}$$

where the final equality uses the fact that $s \geq 3d/4$. We deduce by Proposition 3.14 that

$$\frac{1}{n} \sum_{i=1}^n \{m_{\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i)\}^2 = O_{\mathcal{P}}(\widetilde{K}_Z^{-2s/d_Z} + \widetilde{K}_Z n^{-1}). \tag{3.45}$$

By a similar argument as for the II_n term, but conditioning on $\widehat{f}$ and $\widehat{m}_{\widehat{f}}$ instead of $\widetilde{m}$ and applying (3.45), we conclude that

$$\mathrm{VI}_n = O_{\mathcal{P}}(n^{-\frac{4s}{4s+d}}).$$

We now intend to apply Proposition 3.15 to the VII_n term with $(X,Y,Z) = (\widehat{f}(X,Z), Y, Z)$. By (3.44), we can choose $\sigma_n^2 = \max(\|\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}\|_\infty^2, C^2)$ to fulfil Assumption *(iii)* of that result, and Assumption *(ii)* is satisfied by Assumption 3.5(b). Assumption *(i)* is satisfied with $\zeta_f = \zeta_g = s/d_Z$ by Propositions 3.10 and 3.11, Lemma 3.15, (3.44) and Assumption 3.5(c). We therefore have by Proposition 3.15 that

$$\begin{aligned}\mathrm{VII}_n &= O_{\mathcal{P}}(\widetilde{K}_Z^{-2s/d_Z} + \widetilde{K}_Z^{1/2} n^{-1} + \widetilde{K}_Z^{2-s/d_Z} \log(\widetilde{K}_Z) n^{-2}) \\ &= O_{\mathcal{P}}(n^{-\frac{4s}{4s+d}} + n^{-\frac{4s+d_X}{4s+d}} + \log(n) n^{-\frac{(10s+2d-4d_Z)}{4s+d}}) = O_{\mathcal{P}}(n^{-\frac{4s}{4s+d}})\end{aligned}$$

using that $s \geq 3d/4$. For the final error term, similar to previous terms, for $i \neq j$,

$$\mathbb{E}(\{\widehat{f}(X_i,Z_i) - m_{\widehat{f}}(Z_i)\}\{\widehat{f}(X_j,Z_j) - m_{\widehat{f}}(Z_j)\} \,|\, \widehat{f}, Z_i, Z_j) = 0,$$

so, by Hölder's inequality and the triangle inequality,

$$\begin{aligned}
&\mathbb{E}(|\mathrm{VIII}_n| \,|\, \widehat{f}, \widehat{m}, (Z_i)_{i=1}^n) \\
&\quad \leq \frac{1}{n}\Big(\sum_{i=1}^n \mathbb{E}(\{\widehat{f}(X_i, Z_i) - m_{\widehat{f}}(Z_i)\}^2 \,|\, \widehat{f}, Z_i)\{\widehat{m}(Z_i) - m(Z_i)\}^2\Big)^{1/2} \\
&\quad \leq \frac{2\|\widehat{g}\|_\infty}{n^{1/2}}\Big(\frac{1}{n}\sum_{i=1}^n \{\widehat{m}(Z_i) - m(Z_i)\}^2\Big)^{1/2}.
\end{aligned}$$

Combining Proposition 3.9(b), the argument leading to (3.36), and (3.44) yields that $\|\widehat{g}\|_\infty \leq \|\widehat{\boldsymbol{\beta}}_{XZ}\|_\infty = O_{\mathcal{P}}(1)$. We can therefore apply Corollary 3.2 and Lemma 3.2 as for II_n to conclude that

$$\mathrm{VIII}_n = O_{\mathcal{P}}(n^{-\frac{4s}{4s+d}}).$$

Combining these bounds, we have

$$\frac{1}{n}\sum_{i=1}^n L_i = \tau(1 + R_n),$$

where

$$R_n = O_{\mathcal{P}}(\tau^{-1}n^{-\frac{4s}{4s+d}} + \tau^{-1/2}n^{-1/2}).$$

It follows that

$$R_n = O_{\mathcal{P}_1(\epsilon_n)}(\epsilon_n^{-1}n^{-\frac{4s}{4s+d}} + \epsilon_n^{-1/2}n^{-1/2}),$$

so by Lemma 3.4 and (3.20), (3.40) holds.

To see that (3.41) holds, note that

$$\begin{aligned}
\frac{s_n}{n^{1/2}} \le \Big(\frac{1}{n^2}\sum_{i=1}^n L_i^2\Big)^{1/2} \le 8^{1/2}\Big[&\underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \varepsilon_i^2 f(X_i,Z_i)^2\Big)^{1/2}}_{\widetilde{\mathrm{I}}_n} \\
&+ \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \varepsilon_i^2\{m(Z_i)-\widetilde{m}(Z_i)\}^2\Big)^{1/2}}_{\widetilde{\mathrm{II}}_n} \\
&+ \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n f(X_i,Z_i)^2\{\widehat{g}(X_i,Z_i)-g(X_i,Z_i)\}^2\Big)^{1/2}}_{\widetilde{\mathrm{III}}_n} \\
&+ \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \{Y_i-g(X_i,Z_i)\}^2\{\widehat{g}(X_i,Z_i)-g(X_i,Z_i)\}^2\Big)^{1/2}}_{\widetilde{\mathrm{IV}}_n} \\
&+ \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \varepsilon_i^2 m_{\widehat{f}}(Z_i)^2\Big)^{1/2}}_{\widetilde{\mathrm{V}}_n} + \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \varepsilon_i^2\{m_{\widehat{f}}(Z_i)-\widehat{m}_{\widehat{f}}(Z_i)\}^2\Big)^{1/2}}_{\widetilde{\mathrm{VI}}_n} \\
&+ \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \{m(Z_i)-\widehat{m}(Z_i)\}^2\{m_{\widehat{f}}(Z_i)-\widehat{m}_{\widehat{f}}(Z_i)\}^2\Big)^{1/2}}_{\widetilde{\widetilde{\mathrm{VII}}}_n} \\
&+ \underbrace{\Big(\frac{1}{n^2}\sum_{i=1}^n \{m(Z_i)-\widehat{m}(Z_i)\}^2\{\widehat{f}(X_i,Z_i)-m_{\widehat{f}}(Z_i)\}^2\Big)^{1/2}}_{\widetilde{\widetilde{\mathrm{VIII}}}_n}\Big].
\end{aligned}$$

Combining the bound in (3.42) with Lemma 3.2 yields that $\widetilde{\mathrm{I}}_n = O_{\mathcal{P}}(\tau^{1/2}/n^{1/2})$. All other terms except $\widetilde{\widetilde{\mathrm{VII}}}_n$ are of the same uniform stochastic order as the same expressions for the corresponding terms without tildes. For the final term, then,

$$\widetilde{\widetilde{\mathrm{VII}}}_n \le n^{-1/2}\|\widehat{m}-m\|_\infty\Big(\frac{1}{n}\sum_{i=1}^n\{m_{\widehat{f}}(Z_i)-\widehat{m}_{\widehat{f}}(Z_i)\}^2\Big)^{1/2}.$$

Now $\widehat{m} = \widehat{\boldsymbol{\gamma}}^\top\boldsymbol{\psi}$, and we can let $m^\dagger = \boldsymbol{\gamma}^\top\boldsymbol{\psi}$ denote the $L_2(P)$-best approximant of m over $\mathcal{S}^d_{2r-1,L}$. Then by Proposition 3.9(b), Corollary 3.2, the fact that $s \ge 3d/4$, Propositions 3.10 and 3.11, we have

$$\begin{aligned}
\|\widehat{m}-m\|_\infty \le \|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_\infty + \|m^\dagger - m\|_\infty &= O_{\mathcal{P}}(\widetilde{K}_Z n^{-1/2} + \widetilde{K}_Z^{-s/d_Z}) \\
&= O_{\mathcal{P}}\big(n^{-\frac{2d_X}{4s+d}} + n^{-\frac{2s}{4s+d}}\big) = o_{\mathcal{P}}(1),
\end{aligned}$$

so

$$\widetilde{\mathrm{VII}}_n = o_{\mathcal{P}}(n^{-\frac{4s}{4s+d}})$$

by (3.45). We conclude that $s_n/n^{1/2} = \tau U_n$, where

$$U_n = O_{\mathcal{P}}(\tau^{-1/2}n^{-1/2} + \tau^{-1}n^{-\frac{4s}{4s+d}}) = o_{\mathcal{P}_1(\epsilon_n)}(1),$$

and hence (3.41) is satisfied. This completes the proof.

3.8.8 Proof of Proposition 3.4

Recall that given two probability measures μ and ν on a measurable space $(E, \mathcal{E})$ such that μ has density p with respect to ν, we define the χ^2-divergence from ν to μ by

$$\chi^2(\mu, \nu) := \int_E p^2 \, d\nu - 1.$$

Let $\mathcal{A} := [0,1]^{d_X} \times \{-1, 1\} \times [0,1]^{d_Z}$, and let $P_0 \in \mathcal{P}$ denote a fixed null distribution supported on $\mathcal{A}$. Further, for each $n \in \mathbb{N}$, let $\mathcal{Q}_n \subseteq \mathcal{P}_1(\epsilon_n)$ denote a finite family of alternative distributions supported on $\mathcal{A}$. Suppose that $Q \in \mathcal{Q}_n$ has density $q_Q : \mathcal{A} \to [0, \infty)$ with respect to P_0 and define $P_0^n := P_0^{\otimes n}$ and

$$P_1^n := \frac{1}{|\mathcal{Q}_n|} \sum_{Q \in \mathcal{Q}_n} Q^{\otimes n},$$

where $\otimes n$ denotes the n-fold product of a measure with itself. Suppose that

$$\limsup_{n \to \infty} \chi^2(P_1^n, P_0^n) \leq 1. \tag{3.46}$$

Now, for all $n \in \mathbb{N}$ and tests ϕ,

$$\begin{aligned}\inf_{P \in \mathcal{P}_1(\epsilon_n)} \mathbb{P}_P(\phi = 1) \leq \min_{Q \in \mathcal{Q}_n} \mathbb{P}_Q(\phi = 1) &\leq \frac{1}{|\mathcal{Q}_n|} \sum_{Q \in \mathcal{Q}_n} \mathbb{P}_Q(\phi = 1) \\ &= \int_{\mathcal{A}^n} \phi \, \mathrm{d}P_1^n \leq \mathbb{P}_{P_0}(\phi = 1) + d_{\mathrm{TV}}(P_0^n, P_1^n).\end{aligned}$$

Defining $q_Q^{\otimes n}(x_1, y_1, z_1, \ldots, x_n, y_n, z_n) := \prod_{i=1}^n q_Q(x_i, y_i, z_i)$, we have by Jensen's inequality that

$$\begin{aligned}d_{\mathrm{TV}}(P_0^n, P_1^n)^2 &= \frac{1}{4}\left(\int_{\mathcal{A}^n} \left| \frac{1}{|\mathcal{Q}_n|} \sum_{Q \in \mathcal{Q}_n} q_Q^{\otimes n} - 1 \right| \mathrm{d}P_0^n \right)^2 \\ &\leq \frac{1}{4}\left\{ \int_{\mathcal{A}^n} \left(\frac{1}{|\mathcal{Q}_n|} \sum_{Q \in \mathcal{Q}_n} q_Q^{\otimes n} - 1 \right)^2 \mathrm{d}P_0^n \right\} = \frac{1}{4}\chi^2(P_1^n, P_0^n).\end{aligned}$$

Thus,

$$\limsup_{n \to \infty} \inf_{P \in \mathcal{P}_1(\epsilon_n)} \mathbb{P}_P(\phi = 1) \leq \alpha + \frac{1}{2}$$

for all asymptotically valid tests ϕ, by (3.46).

We now construct P_0 and $\mathcal{Q}_n$ such that (3.46) holds. We let P_0 denote the uniform distribution on $\mathcal{A}$. Then

$$g_{P_0}(x,z) = \mathbb{E}_{P_0}(Y \mid X = x, Z = z) = \mathbb{E}_{P_0}(Y) = 0 = \mathbb{E}_{P_0}(Y \mid Z = z) = m_{P_0}(z),$$

so $\tau_{P_0} = 0$. We also note that g and m are constant functions and so $g, m \in \mathcal{H}_s$ with $\|m\|_{\mathcal{H}_s} = \|g\|_{\mathcal{H}_s} = 0$. It is immediate from similar arguments that the remaining conditions of Assumption 3.5 are satisfied for P_0, so $P_0 \in \mathcal{P}$.

We now aim to construct $\mathcal{Q}_n$. To this end, define the bump function $K : [0,1/2] \to [0,\infty)$ by $K(x) := e^{-\frac{1}{x\cdot(1/2-x)^2}}$, let $I_0 := \left(\int_0^{1/2} K(u)^2\,\mathrm{d}u\right)^{1/2} \in (0,\infty)$ and define $v : \mathbb{R} \to \mathbb{R}$ by $v(x) := \frac{1}{\sqrt{2}I_0} \cdot K(x)\mathbb{1}_{\{x\in[0,1/2]\}} - \frac{1}{\sqrt{2}I_0} \cdot K(x-1/2)\mathbb{1}_{\{x\in[1/2,1]\}}$ and $v(x) := 0$ for $x \in \mathbb{R} \setminus [0,1]$, so that v is infinitely differentiable with $v(0) = v(1) = 0$, $\int_0^1 v(x)\,\mathrm{d}x = 0$ and $\int_0^1 v(x)^2\,\mathrm{d}x = 1$. Now define $h : \mathbb{R}^d \to \mathbb{R}$ by $h(x_1,\dots,x_d) := \prod_{j=1}^d v(x_j)$ and note that h is 0 outside $[0,1]^d$, h is infinitely differentiable, $\int_{\mathbb{R}^d} h^2(x_1,\dots,x_d)\,\mathrm{d}x_1\dots\mathrm{d}x_d = 1$ and $\int_0^1 h(x_1,\dots,x_j,\dots,x_d)\,\mathrm{d}x_j = 0$ for $j \in [d]$.

Define $\rho_n := \lfloor n^{\frac{2}{4s+d}} \rfloor$ and, for $j \in [\rho_n]^d$, define $h_{n,j} : \mathbb{R}^{d_X+d_Z} \to \mathbb{R}$ by $h_{n,j}(x,z) := \rho_n^{d/2} h(\rho_n \cdot (x,z) - j + 1)$, so that $(h_{n,j})_{j\in[\rho_n]^d}$ have disjoint support, $\|h_{n,j}\|_2 = 1$ and $\|h_{n,j}\|_\infty = \rho_n^{d/2}\|h\|_\infty$. Let $\gamma_n := c^{1/2} n^{-\frac{2s+d}{4s+d}}$, where $c \in (0, \rho_n^{-d}\|h\|_\infty^{-2})$ will be specified later. For $\boldsymbol{\eta} := (\eta_j)_{j\in[\rho_n]^d} \in \{-1,1\}^{\rho_n^d}$, define $g_{n,\boldsymbol{\eta}} : \mathbb{R}^{d_X+d_Z} \to (-1,1)$ by

$$g_{n,\boldsymbol{\eta}}(x,z) := \gamma_n \sum_{j\in[\rho_n]^d} \eta_j h_{n,j}(x,z).$$

To see that $g_{n,\boldsymbol{\eta}} \in \mathcal{H}_s^d$, we first note that for any multi-index $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| \le s$, we have

$$\|D^{\boldsymbol{\alpha}} g_{n,\boldsymbol{\eta}}\|_\infty = \gamma_n \rho_n^{d/2+|\boldsymbol{\alpha}|} \|D^{\boldsymbol{\alpha}} h\|_\infty \le c^{1/2} \max_{\widetilde{\boldsymbol{\alpha}}\in\mathbb{N}_0^d : |\widetilde{\boldsymbol{\alpha}}|\le s} \|D^{\widetilde{\boldsymbol{\alpha}}} h\|_\infty. \tag{3.47}$$

Now fix $(x,z),(x',z') \in \mathbb{R}^{d_X+d_Z}$; let $j \in [\rho_n]^d$ denote the unique index such that $h_{n,j}(x,z) \neq 0$ if it exists, and otherwise arbitrarily set $j = (1,\dots,1)^d$. Similarly, let $j' \in [\rho_n]^d$ denote the unique index such that $h_{n,j'}(x',z') \neq 0$ if it exists, and otherwise set $j' = (1,\dots,1)^d$. Then for any $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| = \lceil s \rceil - 1 =: s_0$, we have

$$\begin{aligned}
&|D^{\boldsymbol{\alpha}} g_{n,\boldsymbol{\eta}}(x,z) - D^{\boldsymbol{\alpha}} g_{n,\boldsymbol{\eta}}(x',z')| \\
&\le \gamma_n \rho_n^{d/2+s_0} \big\{ \big| D^{\boldsymbol{\alpha}} h(\rho_n \cdot (x,z) - j + 1) - D^{\boldsymbol{\alpha}} h(\rho_n \cdot (x',z') - j + 1) \big| \\
&\qquad\qquad + \big| D^{\boldsymbol{\alpha}} h(\rho_n \cdot (x,z) - j' + 1) - D^{\boldsymbol{\alpha}} h(\rho_n \cdot (x',z') - j' + 1) \big| \big\} \\
&\le \gamma_n \rho_n^{d/2+s_0} \min\Big(4\|D^{\boldsymbol{\alpha}} h\|_\infty, 2 \max_{\widetilde{\boldsymbol{\alpha}}\in\mathbb{N}_0^d : |\widetilde{\boldsymbol{\alpha}}| = s_0+1} \|D^{\widetilde{\boldsymbol{\alpha}}} h\|_\infty \rho_n \|(x,z) - (x',z')\|_1 \Big) \\
&\le \gamma_n \rho_n^{d/2+s_0} \max\Big(4\|D^{\boldsymbol{\alpha}} h\|_\infty, 2 \max_{\widetilde{\boldsymbol{\alpha}}\in\mathbb{N}_0^d : |\widetilde{\boldsymbol{\alpha}}| \le s_0+1} \|D^{\widetilde{\boldsymbol{\alpha}}} h\|_\infty \Big) \min(1, \rho_n \|(x,z) - (x',z')\|_1) \\
&\le c^{1/2} d^{1/2} \max\Big(4\|D^{\boldsymbol{\alpha}} h\|_\infty, 2 \max_{\widetilde{\boldsymbol{\alpha}}\in\mathbb{N}_0^d : |\widetilde{\boldsymbol{\alpha}}| \le s_0+1} \|D^{\widetilde{\boldsymbol{\alpha}}} h\|_\infty \Big) \|(x,z) - (x',z')\|^{s-s_0},
\end{aligned} \tag{3.48}$$

where the final inequality uses the fact that $\min(1,t)^y \leq t^y$ for any $t > 0$ and $y \in (0,1)$. Using (3.47) and (3.48) and reducing $c > 0$ such that

$$c^{1/2} < \max\Big(4\|D^{\boldsymbol{\alpha}}\|_\infty, 2 \max_{\boldsymbol{\alpha}\in\mathbb{N}_0^d:|\boldsymbol{\alpha}|\leq s_0+1} \|D^{\boldsymbol{\alpha}}\|_\infty, \max_{\boldsymbol{\alpha}\in\mathbb{N}_0^d:|\boldsymbol{\alpha}|\leq s} \|D^{\boldsymbol{\alpha}}h\|_\infty\Big)^{-1} \frac{C}{d^{1/2}},$$

if necessary, we ensure that $g_{n,\boldsymbol{\eta}} \in \mathcal{H}_s^{d_X+d_Z}$ with $\|g_{n,\boldsymbol{\eta}}\|_{\mathcal{H}_s} \leq C$.

Define now $Q_{n,\boldsymbol{\eta}}$ such that (X,Z) is uniform on $[0,1]^{d_X} \times [0,1]^{d_Z}$ and Y is Rademacher with

$$\mathbb{E}_{Q_{n,\boldsymbol{\eta}}}(Y \,|\, X = x, Z = z) = g_{n,\boldsymbol{\eta}}(x,z).$$

Note that by construction

$$m_{n,\boldsymbol{\eta}}(z) := \int_{[0,1]^{d_X}} g_{n,\boldsymbol{\eta}}(x,z)\,\mathrm{d}x = 0 \quad \text{for any } \boldsymbol{\eta} \in \{-1,1\}^{\rho_n^d},$$

so $m_{n,\boldsymbol{\eta}} \in \mathcal{H}_s^{d_Z}$ with $\|m_{n,\boldsymbol{\eta}}\|_{\mathcal{H}_s} = 0$. Further,

$$\tau_{Q_{n,\boldsymbol{\eta}}} = \mathbb{E}_{Q_{n,\boldsymbol{\eta}}}\big[\{g_{n,\boldsymbol{\eta}}(X,Z) - m(Z)\}^2\big] = \gamma_n^2 \rho_n^d \leq c^2 n^{-\frac{4s}{4s+d}},$$

and we deduce from the definition of ϵ_n that $Q_{n,\boldsymbol{\eta}} \in \mathcal{P}_1(\epsilon_n)$ for sufficiently large n. We let $\mathcal{Q}_n := \{Q_{n,\boldsymbol{\eta}} : \boldsymbol{\eta} \in \{-1,1\}^{\rho_n^d}\}$.

To see that (3.46) is satisfied, we note that

$$\begin{aligned}\chi^2(P_1^n, P_0^n) &= -1 + \frac{1}{|\mathcal{Q}_n|^2} \sum_{Q,Q'\in\mathcal{Q}_n} \int_{\mathcal{A}^n} q_Q^{\otimes n} q_{Q'}^{\otimes n}\,\mathrm{d}P_0^n \\ &= -1 + \frac{1}{|\mathcal{Q}_n|^2} \sum_{\boldsymbol{\eta},\boldsymbol{\eta}'\in\{-1,1\}^{\rho_n^d}} \Big(\int_{\mathcal{A}} q_{Q_{n,\boldsymbol{\eta}}} q_{Q_{n,\boldsymbol{\eta}'}}\,\mathrm{d}P_0\Big)^n,\end{aligned}$$

so it suffices to show that lim sup of the second term is at most 2 as $n \to \infty$. But

$$q_{Q_{n,\boldsymbol{\eta}}}(x,y,z) = \{1 + g_{n,\boldsymbol{\eta}}(x,z)\}^{(1+y)/2}\{1 - g_{n,\boldsymbol{\eta}}(x,z)\}^{(1-y)/2},$$

so, for $Q_{n,\boldsymbol{\eta}}, Q_{n,\boldsymbol{\eta}'} \in \mathcal{Q}_n$, we have

$$\begin{aligned}
\int_{\mathcal{A}} q_{Q_{n,\boldsymbol{\eta}}} q_{Q_{n,\boldsymbol{\eta}'}} \,\mathrm{d}P_0 &= \frac{1}{2}\mathbb{E}_{P_0}(q_{Q_{n,\boldsymbol{\eta}}}(X,1,Z) q_{Q_{n,\boldsymbol{\eta}'}}(X,1,Z) \,|\, Y=1) \\
&\qquad + \frac{1}{2}\mathbb{E}_{P_0}(q_{Q_{n,\boldsymbol{\eta}}}(X,-1,Z) q_{Q_{n,\boldsymbol{\eta}'}}(X,-1,Z) \,|\, Y=-1) \\
&= \frac{1}{2}\mathbb{E}_{P_0}(\{1+g_{n,\boldsymbol{\eta}}(X,Z)\}\{1+g_{n,\boldsymbol{\eta}'}(X,Z)\}) \\
&\qquad + \frac{1}{2}\mathbb{E}_{P_0}(\{1-g_{n,\boldsymbol{\eta}}(X,Z)\}\{1-g_{n,\boldsymbol{\eta}'}(X,Z)\}) \\
&= 1 + \mathbb{E}_{P_0}(g_{n,\boldsymbol{\eta}}(X,Z) g_{n,\boldsymbol{\eta}'}(X,Z)) \\
&= 1 + \gamma_n^2 \sum_{j,j' \in [\rho_n]^d} \eta_j \eta'_{j'} \int_{[0,1]^{d_X+d_Z}} h_{n,j}(x,z) h_{n,j'}(x,z) \,\mathrm{d}x \,\mathrm{d}z \\
&= 1 + \gamma_n^2 \boldsymbol{\eta}^\top \boldsymbol{\eta}'.
\end{aligned}$$

Let $\boldsymbol{U} = (U_1, \ldots, U_{\rho_n^d})$ and $\boldsymbol{W} = (W_1, \ldots, W_{\rho_n^d})$ be independent random vectors, each with independent Rademacher components. Then

$$\begin{aligned}
\frac{1}{|\mathcal{Q}_n|^2} \sum_{Q,Q' \in \mathcal{Q}_n} \biggl(\int_{\mathcal{A}} q_Q q_{Q'} \,\mathrm{d}P_0 \biggr)^n &= \frac{1}{2^{2\rho_n^d}} \sum_{\boldsymbol{\eta},\boldsymbol{\eta}' \in \{-1,1\}^{\rho_n^d}} (1 + \gamma_n^2 \boldsymbol{\eta}^\top \boldsymbol{\eta}')^n \\
&\le \frac{1}{2^{2\rho_n^d}} \sum_{\boldsymbol{\eta},\boldsymbol{\eta}' \in \{-1,1\}^{\rho_n^d}} e^{n\gamma_n^2 \boldsymbol{\eta}^\top \boldsymbol{\eta}'} = \mathbb{E}(e^{n\gamma_n^2 \boldsymbol{U}^\top \boldsymbol{W}}) = \prod_{j=1}^{\rho_n^d} \mathbb{E}(e^{n\gamma_n^2 U_j W_j}) \\
&= \cosh(n\gamma_n^2)^{\rho_n^d} \le e^{n^2\gamma_n^4 \rho_n^d/2}.
\end{aligned}$$

Thus,

$$\limsup_{n\to\infty} \frac{1}{|\mathcal{Q}_n|^2} \sum_{Q,Q' \in \mathcal{Q}_n} \biggl(\int_{\mathcal{A}} q_Q q_{Q'} \,\mathrm{d}P_0 \biggr)^n \le \limsup_{n\to\infty} e^{n^2\gamma_n^4 \rho_n^d/2} \le \exp(c^2/2).$$

Taking $c \le \sqrt{2\log 2}$, we have proved (3.46) for P_0 and $\mathcal{Q}_n$, and the result follows.

3.9 Auxiliary lemmas

3.9.1 Uniform convergence results

Recall the 'uniformly small in probability' notation $o_{\mathcal{P}}(1)$ defined in Section 3.1.3. As before, we omit the subscript P in a random variable $X_{P,n}$ and simply write X_n whenever the context is clear.

We will frequently apply a variant of the dominated convergence theorem:

Lemma 3.1. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables. Let $C > 0$ and suppose that $|X_n| \le C$ for all $n \in \mathbb{N}$ and $X_n = o_{\mathcal{P}}(1)$. Then $\sup_{P\in\mathcal{P}} \mathbb{E}_P(|X_n|) = o(1)$.*

Proof. For any given $\epsilon > 0$,

$$|X_n| = |X_n|\mathbb{1}_{\{|X_n|>\epsilon\}} + |X_n|\mathbb{1}_{\{|X_n|\leq\epsilon\}} \leq C\mathbb{1}_{\{|X_n|>\epsilon\}} + \epsilon.$$

By the assumption that $X_n = o_{\mathcal{P}}(1)$, we can choose $N \in \mathbb{N}$ such that $\sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > \epsilon) < \epsilon/C$ for $n \geq N$. It follows that for $n \geq N$,

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P(|X_n|) \leq C \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > \epsilon) + \epsilon < 2\epsilon.$$

Since $\epsilon > 0$ was arbitrary, the result follows. □

The following lemma derives uniform stochastic boundedness of a sequence (X_n) based on a conditional moment condition.

Lemma 3.2. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables on $(\Omega, \mathcal{F})$ and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a sequence of sub-σ-algebras of $\mathcal{F}$. For a positive sequence $(a_n)_{n\in\mathbb{N}}$, possibly depending on P, suppose that $\mathbb{E}_P(|X_n|\,|\,\mathcal{F}_n) = O_{\mathcal{P}}(a_n)$. Then $X_n = O_{\mathcal{P}}(a_n)$.*

Proof. By hypothesis, given $\epsilon > 0$, there exist $M_\epsilon > 0$, $N_\epsilon \in \mathbb{N}$, both depending only on ϵ, such that

$$\sup_{n\geq N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(\mathcal{A}_{n,P}) \leq \frac{\epsilon}{2}, \tag{3.49}$$

where $\mathcal{A}_{n,P} := \{\mathbb{E}_P(|X_n|\,|\,\mathcal{F}_n) \geq M_\epsilon a_n\}$. Then, by Markov's inequality, for any $K_\epsilon > 0$,

$$\begin{aligned}
\sup_{n\geq N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| \geq K_\epsilon a_n) &= \sup_{n\geq N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\frac{|X_n|}{a_n} \wedge K_\epsilon \geq K_\epsilon\right) \\
&\leq \frac{1}{K_\epsilon} \sup_{n\geq N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left(\frac{|X_n|}{a_n} \wedge K_\epsilon\right) \\
&\overset{\text{(i)}}{\leq} \frac{1}{K_\epsilon} \sup_{n\geq N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left(\frac{\mathbb{E}_P(|X_n|\,|\,\mathcal{F}_n)}{a_n} \wedge K_\epsilon\right) \\
&\leq \frac{1}{K_\epsilon} \sup_{n\geq N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left\{\left(\frac{\mathbb{E}_P(|X_n|\,|\,\mathcal{F}_n)}{a_n} \wedge K_\epsilon\right)\mathbb{1}_{\mathcal{A}_{n,P}}\right\} \\
&\qquad + \frac{1}{K_\epsilon} \sup_{n\geq N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left\{\left(\frac{\mathbb{E}_P(|X_n|\,|\,\mathcal{F}_n)}{a_n} \wedge K_\epsilon\right)\mathbb{1}_{\mathcal{A}^c_{n,P}}\right\} \\
&\leq \sup_{n\geq N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(\mathcal{A}_{n,P}) + \frac{M_\epsilon}{K_\epsilon} \overset{\text{(ii)}}{\leq} \frac{\epsilon}{2} + \frac{M_\epsilon}{K_\epsilon},
\end{aligned}$$

where step (i) uses conditional Jensen's inequality and step (ii) uses the inequality (3.49). Then the desired result follows by taking $K_\epsilon \geq 2M_\epsilon/\epsilon$. □

Lemma 3.3. *Let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be sequences of real-valued random variables. If $X_n = o_{\mathcal{P}}(1)$ and $Y_n = O_{\mathcal{P}}(1)$ then $X_nY_n = o_{\mathcal{P}}(1)$.*

Proof. Let $\epsilon > 0$ be given. Then for any $M > 0$

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_nY_n| > \epsilon) \le \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > \epsilon/M) + \sup_{P\in\mathcal{P}} \mathbb{P}_P(|Y_n| > M).$$

Choose $M > 0$ and $n_0 \in \mathbb{N}$ large enough that $\sup_{P\in\mathcal{P}} \mathbb{P}_P(|Y_n| > M) \le \epsilon/2$ for all $n \ge n_0$. By increasing n_0 if necessary, we can ensure that $\sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > \epsilon/M) \le \epsilon/2$ for all $n \ge n_0$. The result follows. □

Lemma 3.4. *Let $(X_n)_{n\in\mathbb{N}}$ and $(R_n)_{n\in\mathbb{N}}$ be sequences of real-valued random variables. Suppose that $R_n > 0$ for all $n \in \mathbb{N}$, $X_n = O_{\mathcal{P}}(R_n)$ and $R_n = o_{\mathcal{P}}(1)$. Then $X_n = o_{\mathcal{P}}(1)$.*

Proof. By hypothesis, for any $\epsilon > 0$, there exist constants $M_\epsilon > 0$ and $N_\epsilon \in \mathbb{N}$, both depending only on ϵ, such that

$$\sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(R_n > \epsilon/M_\epsilon) \le \epsilon/2 \quad \text{and} \quad \sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > R_nM_\epsilon) \le \epsilon/2.$$

Therefore,

$$\begin{aligned}\sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > \epsilon) \le{}& \sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > \epsilon, R_n > \epsilon/M_\epsilon) \\ &+ \sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > \epsilon, R_n \le \epsilon/M_\epsilon) \\ \le{}& \sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(R_n > \epsilon/M_\epsilon) + \sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > R_nM_\epsilon) \le \epsilon,\end{aligned}$$

as required. □

Lemma 3.5. *Let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be sequences of real-valued random variables. For positive sequences $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, suppose that $X_n = O_{\mathcal{P}}(a_n)$ and $Y_n = O_{\mathcal{P}}(b_n)$. Then $X_nY_n = O_{\mathcal{P}}(a_nb_n)$.*

Proof. For any $\epsilon > 0$, there exist $N_\epsilon \in \mathbb{N}$ and $M_\epsilon, K_\epsilon > 0$, all depending only on ϵ, such that

$$\sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > a_nM_\epsilon) \le \epsilon/2 \quad \text{and} \quad \sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|Y_n| > b_nK_\epsilon) \le \epsilon/2.$$

Notice that if $|X_nY_n| > a_nb_nM_\epsilon K_\epsilon$, then either $|X_n| > a_nM_\epsilon$ or $|Y_n| > b_nK_\epsilon$. Therefore, by a union bound,

$$\begin{aligned}\sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_nY_n| > a_nb_nM_\epsilon K_\epsilon) \le{}& \sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > a_nM_\epsilon) \\ &+ \sup_{n\ge N_\epsilon} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|Y_n| > b_nK_\epsilon) \le \epsilon,\end{aligned}$$

as desired. □

Lemma 3.6. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables on $(\Omega, \mathcal{F})$, and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a sequence of sub-σ-algebras of $\mathcal{F}$. Suppose that $|X_n| = Y_n + Z_n$. If $Y_n = o_{\mathcal{P}}(1)$ and $\mathbb{E}_P(Z_n \,|\, \mathcal{F}_n) = o_{\mathcal{P}}(1)$, then $X_n = o_{\mathcal{P}}(1)$.*

Proof. Let $\epsilon \in (0, 1/2]$ be given. By Markov's inequality,

$$\begin{aligned}\sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| > \epsilon) = \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| \wedge \epsilon > \epsilon) &\leq \frac{1}{\epsilon} \sup_{P\in\mathcal{P}} \mathbb{E}_P(|X_n| \wedge \epsilon) \\ &\leq \frac{1}{\epsilon} \sup_{P\in\mathcal{P}} \mathbb{E}_P\big((\epsilon^2 + Z_n) \wedge \epsilon\big) + \sup_{P\in\mathcal{P}} \mathbb{P}_P(|Y_n| > \epsilon^2).\end{aligned}$$

The second term converges to 0 by assumption. For the first term, by Jensen's inequality,

$$\begin{aligned}\frac{1}{\epsilon} \sup_{P\in\mathcal{P}} \mathbb{E}_P\big((\epsilon^2 + Z_n) \wedge \epsilon\big) &= \frac{1}{\epsilon} \sup_{P\in\mathcal{P}} \mathbb{E}_P\Big[\mathbb{E}_P\big((\epsilon^2 + Z_n) \wedge \epsilon \,|\, \mathcal{F}_n\big)\Big] \\ &\leq \frac{1}{\epsilon} \sup_{P\in\mathcal{P}} \mathbb{E}_P\Big[\big\{\epsilon^2 + \mathbb{E}_P(Z_n \,|\, \mathcal{F}_n)\big\} \wedge \epsilon\Big] \\ &\leq 2\epsilon + \sup_{P\in\mathcal{P}} \mathbb{P}_P\big(|\mathbb{E}_P(Z_n \,|\, \mathcal{F}_n)| > \epsilon^2\big).\end{aligned}$$

The result therefore follows by our hypothesis on $\mathbb{E}_P(Z_n \,|\, \mathcal{F}_n)$. □

Lemma 3.7. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables and let X be another such variable. Assume that $|X_n - X| = o_{\mathcal{P}}(1)$ and let $h : \mathbb{R} \to \mathbb{R}$ be a continuous function. Suppose that at least one of the following conditions hold:*

(i) h is uniformly continuous,

(ii) X is uniformly tight, that is,

$$\lim_{M\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X| > M) = 0.$$

Then $|h(X_n) - h(X)| = o_{\mathcal{P}}(1)$.

Proof. Let $\epsilon > 0$ be given. We need to show that

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|h(X_n) - h(X)| > \epsilon) = 0.$$

If h is uniformly continuous, then we can find $\delta > 0$ such that $|h(x) - h(y)| \leq \epsilon$ whenever $|x - y| \leq \delta$. Thus,

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P(|h(X_n) - h(X)| > \epsilon) \leq \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n - X| > \delta) \to 0$$

as $n \to \infty$. On the other hand, suppose now that X is uniformly tight and let $M > 0$. Since h is continuous, it is uniformly continuous on $[-M, M]$, so we can choose $\delta > 0$ such that

$|h(x) - h(y)| \le \epsilon$ whenever $x, y \in [-M, M]$ satisfy $|x - y| \le \delta$. Hence, for $M > \delta$,

$$\begin{aligned}\sup_{P\in\mathcal{P}} \mathbb{P}_P(|h(X_n) - h(X)| > \epsilon) &\le \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n - X| > \delta) \\ &\quad + \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n| \vee |X| > M, |X_n - X| \le \delta) \\ &\le \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X_n - X| > \delta) + \sup_{P\in\mathcal{P}} \mathbb{P}_P(|X| > M - \delta) \to 0\end{aligned}$$

as $n, M \to \infty$. □

Lemma 3.8. *Let $(X_{n,m})_{n\in\mathbb{N},m\in[n]}$ be a triangular array of real-valued random variables and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration on $\mathcal{F}$. Assume that*

(i) *$X_{n,1}, \dots, X_{n,n}$ are conditionally independent given $\mathcal{F}_n$, for each $n \in \mathbb{N}$;*

(ii) *$\mathbb{E}_P(X_{n,m} \,|\, \mathcal{F}_n) = 0$ for all $n \in \mathbb{N}, m \in [n]$;*

(iii) *$|n^{-1} \sum_{m=1}^n \mathbb{E}_P(X_{n,m}^2 \,|\, \mathcal{F}_n) - 1| = o_{\mathcal{P}}(1)$;*

(iv) *there exists $\delta > 0$ such that*

$$\frac{1}{n} \sum_{m=1}^n \mathbb{E}_P(|X_{n,m}|^{2+\delta} \,|\, \mathcal{F}_n) = o_{\mathcal{P}}(n^{\delta/2}).$$

Then $S_n := n^{-1/2} \sum_{m=1}^n X_{n,m}$ converges uniformly in distribution to $N(0,1)$, i.e.

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{x\in\mathbb{R}} |\mathbb{P}_P(S_n \le x) - \Phi(x)| = 0.$$

Proof. We will make the dependence of $X_{n,m}$ and $\mathcal{F}_n$ on P clear by instead writing $X_{P,n,m}$ and $\mathcal{F}_{P,n}$ throughout. By Kasy (2019, Lemma 1) it suffices to show that

$$\frac{1}{\sqrt{n}} \sum_{m=1}^n X_{P_n,n,m} \xrightarrow{d} \mathcal{N}(0,1)$$

for any sequence $(P_n)_{n\in\mathbb{N}}$ in $\mathcal{P}$. Define the triangular array $W_{n,m} := n^{-1/2} X_{P_n,n,m}$ for $n \in \mathbb{N}$ and $m \in [n]$, and let $\tilde{\mathcal{F}}_{n,m}$ be the smallest σ-algebra containing $\mathcal{F}_{P_n,n}$ that makes $X_{P_n,n,1}, \dots, X_{P_n,n,m}$ measurable (and $\tilde{\mathcal{F}}_{n,0} := \mathcal{F}_{P_n,n}$). We claim that $(W_{n,m}, \tilde{\mathcal{F}}_{n,m})$ form a martingale difference array. To see this, observe that $W_{n,m}$ is $\tilde{\mathcal{F}}_{n,m}$-measurable and

$$\begin{aligned}\mathbb{E}_{P_n}(W_{n,m} \,|\, \tilde{\mathcal{F}}_{n,m-1}) &= \frac{1}{n^{1/2}} \mathbb{E}_{P_n}(X_{P_n,n,m} \,|\, \mathcal{F}_{P_n,n}, X_{P_n,n,1}, \dots, X_{P_n,n,m-1}) \\ &= \frac{1}{n^{1/2}} \mathbb{E}_{P_n}(X_{P_n,n,m} \,|\, \mathcal{F}_{P_n,n}) = 0,\end{aligned}$$

where we have used assumptions (i) and (ii) in the penultimate and final equalities, respectively, and this establishes our claim. Now

$$\sum_{m=1}^{n} \mathbb{E}_{P_n}(W_{n,m}^2 \,|\, \tilde{\mathcal{F}}_{n,m-1}) = \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{P_n}(X_{P_n,n,m}^2 \,|\, \mathcal{F}_{P_n,n}) \xrightarrow{P} 1,$$

by assumptions (i) and (iii), and

$$\sum_{m=1}^{n} \mathbb{E}_{P_n}(|W_{n,m}|^{2+\delta} \,|\, \tilde{\mathcal{F}}_{n,m-1}) = \frac{1}{n^{1+\delta/2}} \sum_{m=1}^{n} \mathbb{E}_{P_n}(|X_{P_n,n,m}|^{2+\delta} \,|\, \mathcal{F}_{P_n,n}) \xrightarrow{P} 0,$$

by assumptions (i) and (iv). It follows that for any $c > 0$,

$$\sum_{m=1}^{n} \mathbb{E}_{P_n}(|W_{n,m}|^2 \mathbb{1}_{\{|W_{n,m}|>c\}} \,|\, \tilde{\mathcal{F}}_{n,m-1}) < \frac{1}{c^\delta} \sum_{m=1}^{n} \mathbb{E}_{P_n}(|W_{n,m}|^{2+\delta} \,|\, \tilde{\mathcal{F}}_{n,m-1}) \xrightarrow{P} 0,$$

so the conditional Lindeberg condition is satisfied. The result therefore follows by the Lindeberg–Feller central limit theorem for martingales (e.g. Durrett, 2019, Theorem 8.2.4). □

Lemma 3.9. *Let $(X_{n,m})_{n\in\mathbb{N},m\in[n]}$ be a triangular array of real-valued random variables and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration on $\mathcal{F}$. Assume that*

(i) $X_{n,1}, \ldots, X_{n,n}$ are conditionally independent given $\mathcal{F}_n$ for all $n \in \mathbb{N}$;

(ii) there exists $\delta \in (0, 1]$ such that

$$\sum_{m=1}^{n} \mathbb{E}_P(|X_{n,m}|^{1+\delta} \,|\, \mathcal{F}_n) = o_{\mathcal{P}}(n^{1+\delta}).$$

Then $S_n := n^{-1} \sum_{m=1}^{n} X_{n,m}$ and $\mu_{P,n} := n^{-1} \sum_{m=1}^{n} \mathbb{E}_P(X_{n,m} \,|\, \mathcal{F}_n)$ satisfy $|S_n - \mu_{P,n}| = o_{\mathcal{P}}(1)$; i.e., for any $\epsilon > 0$

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|S_n - \mu_{P,n}| > \epsilon) = 0.$$

Proof. For $n \in \mathbb{N}$, $m \in [n]$, define $W_{n,m} := X_{n,m} - \mu_{P,n}$. Note that

$$\begin{aligned} \sup_{P\in\mathcal{P}} \sum_{m=1}^{n} \mathbb{E}_P(|W_{n,m}|^{1+\delta} \,|\, \mathcal{F}_n) &\leq 2^\delta \left(\sup_{P\in\mathcal{P}} \sum_{m=1}^{n} \mathbb{E}_P(|X_{n,m}|^{1+\delta} \,|\, \mathcal{F}_n) + n|\mu_{P,n}|^{1+\delta} \right) \\ &\leq 2^{\delta+1} \left(\sup_{P\in\mathcal{P}} \sum_{m=1}^{n} \mathbb{E}_P(|X_{n,m}|^{1+\delta} \,|\, \mathcal{F}_n) \right) = o_{\mathcal{P}}(n^{1+\delta}), \end{aligned} \tag{3.50}$$

by assumption (ii). We need to show that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\left| \frac{1}{n} \sum_{m=1}^{n} W_{n,m} \right| \geq \epsilon \right) = 0.$$

Define $W_{n,m}^{<} := W_{n,m}\mathbb{1}_{\{|W_{n,m}|\le n\}}$ and $W_{n,m}^{>} := W_{n,m}\mathbb{1}_{\{|W_{n,m}|>n\}}$. By the triangle inequality we can write

$$\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\Big|\frac{1}{n}\sum_{m=1}^{n}W_{n,m}\Big|\ge\epsilon\Big) \le \underbrace{\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\Big|\frac{1}{n}\sum_{m=1}^{n}[W_{n,m}^{<}-\mathbb{E}_P(W_{n,m}^{<}\,|\,\mathcal{F}_n)]\Big|\ge\frac{\epsilon}{3}\Big)}_{\mathrm{I}_n}$$
$$+\underbrace{\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\Big|\frac{1}{n}\sum_{m=1}^{n}W_{n,m}^{>}\Big|\ge\frac{\epsilon}{3}\Big)}_{\mathrm{II}_n}+\underbrace{\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\Big|\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_P(W_{n,m}^{<}\,|\,\mathcal{F}_n)\Big|\ge\frac{\epsilon}{3}\Big)}_{\mathrm{III}_n},$$

and we will treat each term separately. Considering first I_n, we note that

$$\begin{aligned}\mathrm{I}_n &= \sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\Big|\frac{1}{n}\sum_{m=1}^{n}[W_{n,m}^{<}-\mathbb{E}_P(W_{n,m}^{<}\,|\,\mathcal{F}_n)]\Big|\wedge\frac{\epsilon}{3}\ge\frac{\epsilon}{3}\Big)\\ &\le \frac{9}{\epsilon^2}\sup_{P\in\mathcal{P}}\mathbb{E}_P\Big(\Big\{\frac{1}{n}\sum_{m=1}^{n}[W_{n,m}^{<}-\mathbb{E}_P(W_{n,m}^{<}\,|\,\mathcal{F}_n)]\Big\}^2\wedge\frac{\epsilon^2}{9}\Big)\\ &\le \frac{9}{\epsilon^2}\sup_{P\in\mathcal{P}}\mathbb{E}_P\Big(\mathbb{E}_P\Big[\Big\{\frac{1}{n}\sum_{m=1}^{n}[W_{n,m}^{<}-\mathbb{E}_P(W_{n,m}^{<}\,|\,\mathcal{F}_n)]\Big\}^2\,\Big|\,\mathcal{F}_n\Big]\wedge\frac{\epsilon^2}{9}\Big),\end{aligned}$$

where we have applied Markov's inequality and the tower property combined with the monotonicity of conditional expectations to move the minimum inside the conditional expectation. By assumption (i), the terms in the sum of squares are conditionally independent, so the cross terms vanish, and we find

$$\begin{aligned}\mathrm{I}_n &\le \frac{9}{\epsilon^2}\sup_{P\in\mathcal{P}}\mathbb{E}_P\Big(\Big\{\frac{1}{n^2}\sum_{m=1}^{n}\mathrm{Var}_P(W_{n,m}^{<}\,|\,\mathcal{F}_n)\Big\}\wedge\frac{\epsilon^2}{9}\Big)\\ &\le \frac{9}{\epsilon^2}\sup_{P\in\mathcal{P}}\mathbb{E}_P\Big(\frac{1}{n^2}\sum_{m=1}^{n}\mathbb{E}_P\{(W_{n,m}^{<})^2\,|\,\mathcal{F}_n\}\wedge\frac{\epsilon^2}{9}\Big).\end{aligned}$$

Now, for $\delta\in(0,1)$,

$$\begin{aligned}\mathbb{E}_P\{(W_{n,m}^{<})^2\,|\,\mathcal{F}_n\} &= \int_0^\infty\mathbb{P}_P((W_{n,m}^{<})^2>t\,|\,\mathcal{F}_n)\,\mathrm{d}t = \int_0^\infty 2y\,\mathbb{P}_P(|W_{n,m}^{<}|>y\,|\,\mathcal{F}_n)\,\mathrm{d}y\\ &\le \int_0^n 2y\,\mathbb{P}_P(|W_{n,m}|>y\,|\,\mathcal{F}_n)\,\mathrm{d}y = n^2\int_0^1 2u\,\mathbb{P}_P(|W_{n,m}|>nu\,|\,\mathcal{F}_n)\,\mathrm{d}u\\ &\le n^{1-\delta}\Big(\int_0^1 2u^{-\delta}\,\mathrm{d}u\Big)\mathbb{E}_P(|W_{n,m}|^{1+\delta}\,|\,\mathcal{F}_n) = \frac{2}{1-\delta}n^{1-\delta}\mathbb{E}_P(|W_{n,m}|^{1+\delta}\,|\,\mathcal{F}_n),\end{aligned}$$

where we have used the substitutions $y=\sqrt{t}$ and $u=(1/n)y$, as well as the conditional version of Markov's inequality. We deduce that for any $\delta\in(0,1]$,

$$\mathrm{I}_n\le\frac{9}{\epsilon^2}\Big(\frac{2\mathbb{1}_{\{\delta\in(0,1)\}}}{1-\delta}+\mathbb{1}_{\{\delta=1\}}\Big)\sup_{P\in\mathcal{P}}\mathbb{E}_P\Big(\frac{1}{n^{1+\delta}}\sum_{m=1}^{n}\mathbb{E}_P\{(|W_{n,m}|^{1+\delta}\,|\,\mathcal{F}_n\}\wedge\frac{\epsilon^2}{9}\Big)\to 0,$$

by (3.50) and Lemma 3.1.

We now deal with II_n and III_n by first noting that, using similar $\epsilon/3$-thresholding as above, we obtain

$$\begin{aligned}\mathrm{III}_n &= \sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\Big|\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_P(W^{<}_{n,m}\,|\,\mathcal{F}_n)\Big| \wedge \frac{\epsilon}{3}\Big| \geq \frac{\epsilon}{3}\Big)\\ &\leq \frac{3}{\epsilon}\sup_{P\in\mathcal{P}} \mathbb{E}_P\Big(\Big|\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_P(W^{<}_{n,m}\,|\,\mathcal{F}_n)\Big| \wedge \frac{\epsilon}{3}\Big)\end{aligned}$$

by Markov's inequality. Now, by construction, we can write

$$\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_P(W^{<}_{n,m}\,|\,\mathcal{F}_n) = -\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_P(W^{>}_{n,m}\,|\,\mathcal{F}_n),$$

and thus by the triangle inequality,

$$\mathrm{III}_n \leq \frac{3}{\epsilon}\sup_{P\in\mathcal{P}} \mathbb{E}_P\Big(\Big[\frac{1}{n}\sum_{m=1}^{n}\big|\mathbb{E}_P(W^{>}_{n,m}\,|\,\mathcal{F}_n)\big|\Big] \wedge \frac{\epsilon}{3}\Big) + \frac{3}{\epsilon}\sup_{P\in\mathcal{P}} \mathbb{E}_P\Big(|R_n| \wedge \frac{\epsilon}{3}\Big).$$

The second term converges to 0 by Lemma 3.1, so it remains to show that the first term converges to 0. Now II_n can be seen to also be upper bounded by the first term by a similar argument to the one given above, so we are done if we can show that the first term converges to 0. Applying conditional Hölder's inequality (Gut, 2013, Theorem 10.1.6) followed by conditional Markov's inequality yields

$$\begin{aligned}&\frac{3}{\epsilon}\sup_{P\in\mathcal{P}}\mathbb{E}_P\Big(\Big[\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_P(|W^{>}_{n,m}|\,|\,\mathcal{F}_n)\Big] \wedge \frac{\epsilon}{3}\Big)\\ &\quad\leq \frac{3}{\epsilon}\sup_{P\in\mathcal{P}} \mathbb{E}_P\Big(\Big[\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_P(|W_{n,m}|^{1+\delta}\,|\,\mathcal{F}_n)^{\frac{1}{1+\delta}}\mathbb{P}_P(|W_{n,m}| > n\,|\,\mathcal{F}_n)^{\frac{\delta}{1+\delta}}\Big] \wedge \frac{\epsilon}{3}\Big)\\ &\quad\leq \frac{3}{\epsilon}\sup_{P\in\mathcal{P}} \mathbb{E}_P\Big(\Big[\frac{1}{n^{1+\delta}}\sum_{m=1}^{n}\mathbb{E}_P(|W_{n,m}|^{1+\delta}\,|\,\mathcal{F}_n)\Big] \wedge \frac{\epsilon}{3}\Big).\end{aligned}$$

Finally, combining the above with (3.50) and Lemma 3.1 yields the desired result. □

Lemma 3.10. *Let $\mathcal{P}$ denote a family of distributions of (Y, Z) taking values in $\mathbb{R}\times\mathbb{R}^d$. Define $\boldsymbol{\Sigma_P} := \mathbb{E}_P(ZZ^T)$ and suppose this is invertible for all $P \in \mathcal{P}$. Let $\boldsymbol{\beta}_P := \boldsymbol{\Sigma}_P^{-1}\mathbb{E}_P(ZY)$, $\varepsilon_P := Y - \boldsymbol{\beta}_P^\top Z$ and $\boldsymbol{\Theta}_P := \mathbb{E}(ZZ^T\varepsilon^2)$. Suppose there exist $C, c, \delta > 0$ such that the following hold.*

(i) $\inf_{P\in\mathcal{P}} \min\{\lambda_{\min}(\boldsymbol{\Theta}_P), \lambda_{\min}(\boldsymbol{\Sigma}_P)\} \geq c$.

(ii) $\sup_{P\in\mathcal{P}} \max\{\mathbb{E}_P(\|Z\varepsilon\|_2^{2+\delta}), \mathbb{E}_P(\|Z\|_2^{2+\delta})\} \leq C$.

Given n independent copies $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ of (Y, Z), let $\widehat{\boldsymbol{\beta}}$ denote the ordinary least squares estimator. Then,

$$\sqrt{n}\boldsymbol{\Theta}_P^{-1/2}\boldsymbol{\Sigma}_P(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_P)$$

converges uniformly to a standard d-variate Gaussian distribution.

Proof. Let $\widehat{\mathbf{\Sigma}} := n^{-1}\sum_{i=1}^n Z_i Z_i^\top$. We first argue that

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}} = o_{\mathcal{P}}(1). \tag{3.51}$$

By the equivalence of finite-dimensional norms, it suffices to show that $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_\infty = o_{\mathcal{P}}(1)$, which is equivalent to $\widehat{\mathbf{\Sigma}}_{jk} - \mathbf{\Sigma}_{P,jk} = o_{\mathcal{P}}(1)$ for all $j,k \in [d]$. To show this final claim, let us fix $j,k \in [d]$, and note that

$$(\widehat{\mathbf{\Sigma}})_{jk} - (\mathbf{\Sigma}_P)_{jk} = \frac{1}{n}\sum_{i=1}^n \{(Z_i)_j(Z_i)_k - \mathbb{E}_P((Z)_j(Z)_k)\}.$$

Define the triangular array $X_{n,i} := (Z_i)_j(Z_i)_k - \mathbb{E}_P((Z)_j(Z)_k)$ for $i \in [n]$ and $n \in \mathbb{N}$. Our claim now follows from applying Lemma 3.9 to $X_{n,i}$ where we condition on the trivial σ-algebra and set $\mu_n = 0$. Indeed, (i) and (ii) of Lemma 3.9 are clearly satisfied, and for the third condition we have by the Cauchy–Schwarz inequality,

$$\frac{1}{n}\sum_{m=1}^n \mathbb{E}_P(|X_{n,m}|^{1+\delta/2}) \le 2^{1+\delta/2}\mathbb{E}_P(|(Z)_j(Z)_k|^{1+\delta/2}) \le 2^{1+\delta/2}\mathbb{E}_P(\|Z\|_\infty^{2+\delta}) \le 2^{1+\delta/2}C.$$

Thus, (3.51) follows.

We now argue that

$$\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}} = o_{\mathcal{P}}(1). \tag{3.52}$$

By Weyl's inequality and our assumption on $\lambda_{\min}(\mathbf{\Sigma}_P)$, we have

$$\lambda_{\min}(\widehat{\mathbf{\Sigma}}) \ge \lambda_{\min}(\mathbf{\Sigma}_P) - \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}} \ge c - \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}},$$

thus on the event $\{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}} \le c/2\}$ we have that $\widehat{\mathbf{\Sigma}}$ is invertible. On this event, we have

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}} &= \|\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{\Sigma}_P - \widehat{\mathbf{\Sigma}})\mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}} \\ &\le (\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}} + \|\mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}})\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}}\|\mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}}\end{aligned}$$

and furthermore since $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}}\|\mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}} \le 1/2$ on the event, we deduce that

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}} &\le \frac{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}}\|\mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}}^2}{1 - \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}}\|\mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}}} \\ &\le 2\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}}\|\mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}}^2 \le 2c^{-2}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}}.\end{aligned}$$

Putting things together, for any $\epsilon > 0$, we have

$$\mathbb{P}_P(\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}} \ge \epsilon) \le \mathbb{P}_P(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}} \ge c/2) + \mathbb{P}_P(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_P\|_{\mathrm{op}} \ge c^2\epsilon/2).$$

Thus taking suprema over $\mathcal{P}$ and applying (3.51), we have shown (3.52).

We now turn to proving the stated result. Defining $U_n := n^{-1/2} \sum_{i=1}^n \boldsymbol{\Theta}_P^{-1/2} Z_i \varepsilon_i$, we have

$$\sqrt{n} \boldsymbol{\Theta}_P^{-1/2} \boldsymbol{\Sigma}_P (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_P) = \boldsymbol{\Theta}_P^{-1/2} \boldsymbol{\Sigma}_P \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Theta}_P^{1/2} U_n.$$

We have that the summands in the definition of U_n are mean zero with covariance equal to the identity matrix, and they satisfy Lyapunov's condition since

$$\mathbb{E}(\|\boldsymbol{\Theta}_P^{-1/2} Z \varepsilon\|_2^{2+\delta}) \leq \lambda_{\min}(\boldsymbol{\Theta}_P)^{-(1/2+\delta/4)} \mathbb{E}(\|Z\varepsilon\|_2^{2+\delta}) \leq c^{-(1/2+\delta/4)} C.$$

The Lindeberg–Feller theorem (Vaart, 1998, Proposition 2.27) therefore yields that U_n converges uniformly to a d-variate standard Gaussian. Combining this with a uniform version of Slutsky's theorem (Bengs and Holzmann, 2019, Theorem 6.3) and (3.52) yields the desired result. □

3.9.2 Miscellaneous results

Proposition 3.5. *Let X, Y, Z be random variables with $Y \in \mathbb{R}$, $\mathrm{Var}(Y \mid X, Z) > 0$ almost surely and $\mathbb{E}(Y^4) < \infty$. Then*

$$\frac{(\mathbb{E}[\{Y - \mathbb{E}(Y \mid Z)\} f(X, Z)])^2}{\mathbb{E}[\{Y - \mathbb{E}(Y \mid X, Z)\}^2 f(X, Z)^2]} \tag{3.53}$$

is maximised over f with $0 < \mathbb{E}\{f(X, Z)^4\} < \infty$ by

$$f(X, Z) \propto \frac{\mathbb{E}(Y \mid X, Z) - \mathbb{E}(Y \mid Z)}{\mathrm{Var}(Y \mid X, Z)},$$

and up to positive scaling this is almost surely the unique minimiser.

Proof. Note that the denominator of (3.53) may be written as $\mathbb{E}\{\mathrm{Var}(Y \mid X, Z) f(X, Z)^2\}$. Turning to the numerator, we have by the Cauchy–Schwarz inequality that

$$\begin{aligned}
(\mathbb{E}[\{Y - \mathbb{E}(Y \mid Z)\} f(X, Z)])^2 &= (\mathbb{E}[\{\mathbb{E}(Y \mid X, Z) - \mathbb{E}(Y \mid Z)\} f(X, Z)])^2 \\
&= \left(\mathbb{E}\left[\frac{\sqrt{\mathrm{Var}(Y \mid X, Z)}}{\sqrt{\mathrm{Var}(Y \mid X, Z)}} \{\mathbb{E}(Y \mid X, Z) - \mathbb{E}(Y \mid Z)\} f(X, Z) \right] \right)^2 \\
&\leq \mathbb{E}\{\mathrm{Var}(Y \mid X, Z) f(X, Z)^2\} \, \mathbb{E}\left[\frac{\{\mathbb{E}(Y \mid X, Z) - \mathbb{E}(Y \mid Z)\}^2}{\mathrm{Var}(Y \mid X, Z)} \right],
\end{aligned}$$

with equality if and only if $f(X, Z) \propto \{\mathbb{E}(Y \mid X, Z) - \mathbb{E}(Y \mid Z)\} / \mathrm{Var}(Y \mid X, Z)$ almost surely. □

Proposition 3.6. *Consider the setting of Theorem 3.1. Assume that Assumption 3.3 holds.*

(i) If $Y \perp\!\!\!\perp X \mid Z$, then (3.22) is satisfied.

(ii) If $\widehat{m}$ is formed using a sample independent of $\mathcal{D}_1$, then (3.22) is satisfied.

(iii) If $\widehat{m}$ is a linear smoother, then (3.22) is satisfied.

Proof. As in the main proofs in Appendix 3.8, we suppress dependence on P in what follows.

(i) Under conditional independence, it is immediate that $R_{ij} = 0$.

(ii) Define $w(z_1, z_2) := \mathbb{E}(\{\widehat{m}(z_1) - m(z_1)\}\{\widehat{m}(z_2) - m(z_2)\})$ and note that

$$\mathbb{E}(M_i M_j \,|\, (X_i, Z_i)_{i=1}^n) = w(Z_i, Z_j) = \mathbb{E}(M_i M_j \,|\, (Z_i)_{i=1}^n)$$

by Durrett (2019, Example 5.1.5) since $(X_i, Z_i)_{i=1}^n \perp\!\!\!\perp \widehat{m}$. Thus, $R_{ij} = 0$ and (3.22) is satisfied.

(iii) It suffices to show that $\mathbb{E}(R_{ij}\xi_i\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f}) = 0$. When $\widehat{m}(\cdot) = \sum_{k=1}^n \omega(Z_k, \cdot)Y_k$ is a linear smoother, we note that

$$\begin{aligned}\mathbb{E}(\widehat{m}(Z_i) \,|\, (X_i, Z_i)_{i=1}^n) &= \sum_{k=1}^n \omega(Z_k, Z_i)\mathbb{E}(Y_k \,|\, (X_i, Z_i)_{i=1}^n) \\ &= \sum_{k=1}^n \omega(Z_k, Z_i)\mathbb{E}(Y_k \,|\, (Z_i)_{i=1}^n) \\ &= \mathbb{E}(\widehat{m}(Z_i) \,|\, (Z_i)_{i=1}^n).\end{aligned}$$

Based on this identity, it can be seen that

$$\begin{aligned}R_{ij} &= \mathbb{E}\{\widehat{m}(Z_i)\widehat{m}(Z_j) \,|\, (X_i)_{i=1}^n, (Z_i)_{i=1}^n\} - \mathbb{E}\{\widehat{m}(Z_i)\widehat{m}(Z_j) \,|\, (Z_i)_{i=1}^n\} \\ &= \sum_{k=1}^n \omega(Z_k, Z_i)\omega(Z_k, Z_j)\{\mathbb{E}(Y_k^2 \,|\, (X_i)_{i=1}^n, (Z_i)_{i=1}^n) - \mathbb{E}(Y_k^2 \,|\, (Z_i)_{i=1}^n)\} \\ &\quad + \sum_{1 \le k \ne k' \le n} \omega(Z_k, Z_i)\omega(Z_{k'}, Z_j)\{\mathbb{E}(Y_k Y_{k'} \,|\, (X_i)_{i=1}^n, (Z_i)_{i=1}^n) - \mathbb{E}(Y_k Y_{k'} \,|\, (Z_i)_{i=1}^n)\} \\ &= \sum_{k=1}^n \omega(Z_k, Z_i)\omega(Z_k, Z_j)\{\mathbb{E}(Y_k^2 \,|\, X_k, Z_k) - \mathbb{E}(Y_k^2 \,|\, Z_k)\}.\end{aligned}$$

This identity yields

$$\begin{aligned}&\mathbb{E}(R_{ij}\xi_i\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f}) \\ &\qquad = \sum_{k=1}^n \omega(Z_k, Z_i)\omega(Z_k, Z_j)\mathbb{E}\big[\{\mathbb{E}(Y_k^2 \,|\, X_k, Z_k) - \mathbb{E}(Y_k^2 \,|\, Z_k)\}\xi_i\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f}\big].\end{aligned}$$

When $i \ne j$ at least one of i and j differs from $k \in [n]$. Without loss of generality, assume $k \ne i$ and see

$$\begin{aligned}&\mathbb{E}\Big[\{\mathbb{E}(Y_k^2 \,|\, X_k, Z_k) - \mathbb{E}(Y_k^2 \,|\, Z_k)\}\xi_i\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f}\Big] \\ &= \mathbb{E}\Big[\{\mathbb{E}(Y_k^2 \,|\, X_k, Z_k) - \mathbb{E}(Y_k^2 \,|\, Z_k)\}\underbrace{\mathbb{E}(\xi_i \,|\, Z_i, \widehat{f})}_{=0}\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f}\Big] = 0.\end{aligned}$$

Therefore, we conclude that $\mathbb{E}(R_{ij}\xi_i\xi_j \,|\, (Z_i)_{i=1}^n, \widehat{f}) = 0$ when $\widehat{m}$ is a linear smoother and thus (3.22) is satisfied. □

Lemma 3.11. *Consider the setting of Theorem 3.3 and assume that X and Z are independent for all $P \in \mathcal{P}_0$. Then (3.16) is satisfied for sufficiently small $c > 0$.*

Proof. Recalling the definitions of V and $\mathbf{1} \in \mathbb{R}^{K_X}$ from Proposition 3.13, we note that for every $\boldsymbol{v} \in \mathbb{R}^{K_Z}$, we can define $\boldsymbol{w} := \mathbf{1} \otimes \boldsymbol{v} = (\boldsymbol{I}_{K_Z} \otimes_{\mathrm{Kron}} \mathbf{1})\boldsymbol{v} \in V$ so that

$$\boldsymbol{w}^\top \boldsymbol{\phi}(X, Z) = \boldsymbol{v}^\top \boldsymbol{\phi}^Z(Z),$$

by Proposition 3.13. Therefore,

$$\boldsymbol{w}^\top \boldsymbol{\Lambda}_P \boldsymbol{w} = \mathbb{E}_P(\mathrm{Var}_P(\boldsymbol{v}^\top \boldsymbol{\phi}^Z(Z) \,|\, Z)) = 0.$$

Since $\{x \in \mathbb{R}^{K_{XZ}} : \boldsymbol{\Pi} x = x\} = V^\perp$ is $(K_{XZ} - K_Z)$-dimensional and $\boldsymbol{\Lambda}_P$ is non-negative definite, we conclude that

$$\widetilde{\lambda}_{\min}(\boldsymbol{\Lambda}_P) = \lambda_{K_{XZ}-K_Z-1}(\boldsymbol{\Lambda}_P),$$

where $\lambda_k(\boldsymbol{\Lambda}_P)$ denotes the kth largest eigenvalue of $\boldsymbol{\Lambda}_P$. We can write

$$\boldsymbol{\Lambda}_P = \underbrace{\mathbb{E}_P(\boldsymbol{\phi}(X, Z)\boldsymbol{\phi}(X, Z)^\top)}_{\boldsymbol{\Sigma}_P} - \underbrace{\mathbb{E}_P(\mathbb{E}_P(\boldsymbol{\phi}(X, Z) \,|\, Z)\mathbb{E}(\boldsymbol{\phi}(X, Z) \,|\, Z)^\top)}_{\boldsymbol{\Gamma}_P}.$$

Denote the Kronecker product by $\otimes_{\mathrm{Kron}}$ and note that for $\boldsymbol{x} \in \mathbb{R}^{K_X}$ and $\boldsymbol{z} \in \mathbb{R}^{K_Z}$, we have

$$\boldsymbol{x} \otimes \boldsymbol{z} = (\boldsymbol{I}_{K_Z} \otimes_{\mathrm{Kron}} \boldsymbol{x})\boldsymbol{z}.$$

Write $\boldsymbol{A}_P := \boldsymbol{I}_{K_Z} \otimes_{\mathrm{Kron}} \mathbb{E}_P(\boldsymbol{\phi}^X(X)) \in \mathbb{R}^{K_X K_Z \times K_Z}$. Then, since X and Z are independent,

$$\mathbb{E}_P(\boldsymbol{\phi}(X, Z) \,|\, Z) = \mathbb{E}_P(\boldsymbol{\phi}^X(X)) \otimes \boldsymbol{\phi}^Z(Z) = \boldsymbol{A}_P \boldsymbol{\phi}^Z(Z).$$

Defining $\boldsymbol{\Sigma}_{Z,P} := \mathbb{E}_P(\boldsymbol{\phi}^Z(Z)\boldsymbol{\phi}^Z(Z)^\top) \in \mathbb{R}^{K_Z \times K_Z}$, it follows that $\boldsymbol{\Gamma}_P = \boldsymbol{A}_P \boldsymbol{\Sigma}_{Z,P} \boldsymbol{A}_P^\top$, so we deduce that $\mathrm{rank}(\boldsymbol{\Gamma}_P) \leq \mathrm{rank}(\boldsymbol{\Sigma}_{Z,P}) \leq K_Z$. Hence, by Weyl's inequality,

$$\begin{aligned}\lambda_{K_{XZ}-K_Z-1}(\boldsymbol{\Lambda}_P) &\geq \lambda_{K_{XZ}}(\boldsymbol{\Sigma}_P) + \lambda_{K_{XZ}-K_Z-1}(-\boldsymbol{\Gamma}_P) \\ &= \lambda_{K_{XZ}}(\boldsymbol{\Sigma}_P) - \lambda_{K_Z+1}(\boldsymbol{\Gamma}_P) = \lambda_{K_{XZ}}(\boldsymbol{\Sigma}_P) \geq c_s(r)^d K_{XZ}^{-1} \inf_{(x,z)\in[0,1]^d} p_P(x, z),\end{aligned}$$

by Proposition 3.9(d). This proves the desired claim. □

Corollary 3.1. *Consider the setting of Proposition 3.17. Assume that $\beta = \lambda n^{-1/2}$ and denote $r := n_2/n_1$. Then, given any $\delta > 0$, we can choose $\lambda_0 \equiv \lambda_0(\alpha, \delta, \sigma_{X|Z}, \sigma_{XY|Z}) > 0$ and $r_0 \equiv r_0(\lambda_0, \delta, \sigma_\beta) > 0$ such that*

$$\psi < \frac{1}{2} + \delta,$$

for all $\lambda \geq \lambda_0$ and $r \in (0, r_0]$.

Further, given any $\delta > 0$*, we can choose* $\lambda_1 \equiv \lambda_1(\alpha, \delta, \sigma_{X|Z}, \sigma_{XY|Z}) > 0$ *and* $r_1 \equiv r_1(\lambda_1, \delta, \sigma_\beta) > 0$ *such that*

$$\psi < \alpha + \delta$$

for all $\lambda \in (0, \lambda_1]$ *and* $r \geq r_1$*.*

Proof. To prove the first claim, note that, for $r < 1/2$,

$$\psi \leq \Phi\left(\frac{\lambda r^{1/2}}{\sigma_\beta}\right) + \Phi\left(z_\alpha - \frac{\lambda \sigma_{X|Z}^2}{\sqrt{2}\sigma_{XY|Z}}\right).$$

We can now choose $\lambda_0 \equiv \lambda_0(\alpha, \delta, \sigma_{X|Z}, \sigma_{XY|Z}) > 0$ large enough that the second term is at most $\delta/2$ for $\lambda \geq \lambda_0$ and then choose $r_0 \equiv r_0(\lambda_0, \delta, \sigma_\beta) > 0$ small enough that the first term is less than $1/2 + \delta/2$ for $r \in (0, r_0]$.

To prove the second claim, note that

$$\psi \leq \Phi\left(z_\alpha + \frac{\lambda \sigma_{X|Z}^2}{\sigma_{XY|Z}}\right) + \Phi\left(-\frac{\lambda r^{1/2}}{\sigma_\beta}\right).$$

We can now choose $\lambda_1 \equiv \lambda_1(\alpha, \delta, \sigma_{X|Z}, \sigma_{XY|Z}) > 0$ small enough that the first term is at most $\delta/2 + \alpha$ for $\lambda \in (0, \lambda_1]$ and then choose $r_1 \equiv r_1(\lambda_1, \delta, \sigma_\beta) > 0$ large enough that the second term is less than $\delta/2$ for $r \geq r_1$. □

A discussion of the test proposed by Williamson et al. (2022)

Like our proposal, the test proposed by Williamson et al. (2022) relies on sample splitting. However, their test suffers from a couple of issues as we describe below. To this end, we start by formalising their testing procedure. First split the data $\mathcal{D} = \{(X_i, Y_i, Z_i)\}_{i=1}^{2n}$ randomly into $\mathcal{D}_1$ and $\mathcal{D}_2$ both of size n and let $\mathcal{I}_1$ and $\mathcal{I}_2$ denote the corresponding data indices. We write the sample mean of Y for each split as $\overline{Y}_1 := n^{-1} \sum_{i \in \mathcal{I}_1} Y_i$ and $\overline{Y}_2 := n^{-1} \sum_{i \in \mathcal{I}_2} Y_i$ respectively. Recall the definitions of g and m from Section 3.2.1 and let $\widehat{g}$ and $\widehat{m}$ denote generic estimators of these, where $\widehat{g}$ is constructed on $\mathcal{D}_1$ and $\widehat{m}$ is constructed on $\mathcal{D}_2$. For notational convenience, we define $\mu_0 := \mathbb{E}_P(Y)$, $\sigma_Y^2 := \mathrm{Var}_P(Y)$, $\tau_{xz,0} := \mathbb{E}_P\{(g(X, Z) - \mu_0)^2\}$ and $\tau_{z,0} := \mathbb{E}_P\{(m(Z) - \mu_0)^2\}$. Let us further define

$$\widehat{v}_1 = \frac{\frac{1}{n}\sum_{i \in \mathcal{I}_1}(Y_i - \overline{Y}_1)^2 - \frac{1}{n}\sum_{i \in \mathcal{I}_1}\{Y_i - \widehat{g}(X_i, Z_i)\}^2}{\frac{1}{n}\sum_{i \in \mathcal{I}_1}(Y_i - \overline{Y}_1)^2},$$
$$\widehat{v}_2 = \frac{\frac{1}{n}\sum_{i \in \mathcal{I}_2}(Y_i - \overline{Y}_2)^2 - \frac{1}{n}\sum_{i \in \mathcal{I}_2}(Y_i - \widehat{m}(Z_i))^2}{\frac{1}{n}\sum_{i \in \mathcal{I}_2}(Y_i - \overline{Y}_2)^2},$$

and denote their population counterparts by

$$v_1 := \frac{\tau_{xz,0}}{\sigma_Y^2} \quad \text{and} \quad v_2 := \frac{\tau_{z,0}}{\sigma_Y^2}.$$

As shown in Williamson et al. (2021, Lemma 1), the influence functions of v_1 and v_2 are given by

$$\varphi_1(x,y,z) := \frac{2\{y-g(x,z)\}\{g(x,z)-\mu_0\}+\{g(x,z)-\mu_0\}^2}{\sigma_Y^2} - \tau_{xz,0}\left\{\frac{y-\mathbb{E}_P(Y)}{\sigma_Y^2}\right\}^2$$

and

$$\varphi_2(y,z) := \frac{2\{y-m(z)\}\{m(z)-\mu_0\}+\{m(z)-\mu_0\}^2}{\sigma_Y^2} - \tau_{z,0}\left\{\frac{y-\mathbb{E}_P(Y)}{\sigma_Y^2}\right\}^2$$

respectively. Finally, by letting $\widehat{\eta}_1$ and $\widehat{\eta}_2$ be consistent estimators of $\eta_1 := \mathbb{E}_P(\varphi_1(X,Y,Z)^2)$ and $\eta_2 := \mathbb{E}_P(\varphi_2(Y,Z)^2)$, the test statistic proposed by Williamson et al. (2022) is given as

$$T_{\mathrm{W}} := \frac{\widehat{v}_1 - \widehat{v}_2}{\sqrt{n^{-1}(\widehat{\eta}_1\widehat{\eta}_2)}}.$$

The test statistic T_{W} is calibrated based on a normal approximation and the null of $\tau_P = 0$ is rejected if $T_{\mathrm{W}} > z_{1-\alpha}$. Having specified the test function, we are now ready to describe issues.

1. Lack of power: We begin with a power issue. In particular, we shall see that their test has the asymptotic power equal to its size whenever $\sqrt{n}\tau_P \to 0$, under some regularity conditions. This property is true even for the simple linear model where the optimal detection boundary is known to be $\tau_P \asymp n^{-1}$. To see this, suppose that the assumptions for Williamson et al. (2021, Theorem 1) are satisfied for $\widehat{v}_1$ and $\widehat{v}_2$. That is, $\widehat{v}_1$ and $\widehat{v}_2$ are asymptotically linear with influence functions φ_1 and φ_2, respectively, so that

$$\widehat{v}_1 - v_1 = \frac{1}{n}\sum_{i\in\mathcal{I}_1}\varphi_1(X_i,Y_i,Z_i) + o_P(n^{-1/2})$$

and

$$\widehat{v}_2 - v_2 = \frac{1}{n}\sum_{i\in\mathcal{I}_2}\varphi_2(X_i,Y_i,Z_i) + o_P(n_2^{-1/2}).$$

The asymptotic validity of the approach of Williamson et al. (2022) comes from the fact that the individual influence functions φ_1 and φ_2 are not necessarily degenerate under the null. In particular, when η_1 and η_2 are non-zero, the central limit theorem guarantees that T_{W} converges in distribution to $N(0,1)$ under the null (where $v_1 = v_2$). Similarly, we can also establish the asymptotic normality of T_{W} under the alternative in the case where η_1 and η_2 are non-zero. This asymptotic normality allows us to describe the asymptotic power expression of the given test. More formally, the central limit theorem yields

$$\sqrt{n}(\widehat{v}_1 - v_1) \xrightarrow{d} N(0,\eta_1) \quad \text{and} \quad \sqrt{n}(\widehat{v}_2 - v_2) \xrightarrow{d} N(0,\eta_2).$$

Hence, by Slutsky's theorem and the independence of $\widehat{v}_1$ and $\widehat{v}_2$, we conclude that

$$\frac{(\widehat{v}_1 - \widehat{v}_2) - (v_1 - v_2)}{\sqrt{n^{-1}(\widehat{\eta}_1 + \widehat{\eta}_2)}} \xrightarrow{d} N(0,1).$$

This shows that

$$\mathbb{P}_P(T_{\mathrm{W}} > z_{1-\alpha}) \to \Phi\bigg(z_\alpha + \frac{\sqrt{n}\tau_P}{\sigma_Y^2\sqrt{\eta_1+\eta_2}}\bigg),$$

where we have used the fact that $v_1 - v_2 = \tau_P/\sigma_Y^2$. Therefore, when σ_Y^2, η_1 and η_2 are strictly bounded below by some positive constant, the power converges to the nominal level α whenever $\sqrt{n}\tau_P \to 0$.

2. Asymptotic validity: Another issue concerns asymptotic validity. The previous argument hinges on the condition that η_1 and η_2 are non-zero. As acknowledged by Williamson et al. (2022), the asymptotic validity of their test is no longer guaranteed when η_1 and η_2 are zero. We illustrate this by considering a specific example.

Consider a simple linear model where $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$ and $(X, Z, \varepsilon)^\top$ follows a multivariate normal distribution with zero mean and identity covariance matrix. Assume that $\beta_1 = \beta_2 = 0$. In this scenario, φ_1 and φ_2 are the zero functions on their respective domains, so $\eta_1 = \eta_2 = 0$. Letting $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_1)$ denote the least squares estimator of $(\beta_0, \beta_1, \beta_2)$ based on $\mathcal{D}_1$, and letting F_{k_1,k_2} denote the F-distribution with (k_1, k_2) degrees of freedom, we have

$$\widehat{v}_1 = \frac{\frac{1}{n}\sum_{i\in\mathcal{I}_1}(Y_i - \overline{Y}_1)^2 - \frac{1}{n}\sum_{i\in\mathcal{I}_1}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i - \widehat{\beta}_2 Z_i)^2}{\frac{1}{n}\sum_{i\in\mathcal{I}_1}(Y_i - \overline{Y}_1)^2} \sim \frac{2}{n-1} F_{2,n-1}.$$

Similarly, writing $(\widetilde{\beta}_0, \widetilde{\beta}_2)$ for the least squares estimator of (β_0, β_2) based on $\mathcal{D}_2$, we have

$$\widehat{v}_2 = \frac{\frac{1}{n}\sum_{i\in\mathcal{I}_2}(Y_i - \overline{Y}_2)^2 - \frac{1}{n}\sum_{i\in\mathcal{I}_2}(Y_i - \widetilde{\beta}_0 - \widetilde{\beta}_2 Z_i)^2}{\frac{1}{n}\sum_{i\in\mathcal{I}_2}(Y_i - \overline{Y}_2)^2} \sim \frac{1}{n-1} F_{1,n-1}.$$

Since $F_{2,n-1} \xrightarrow{d} \chi_2^2/2$ and $F_{1,n-1} \xrightarrow{d} \chi_1^2$ where χ_k^2 denotes the chi-square distribution with k degrees of freedom, we observe that

$$n(\widehat{v}_1 - \widehat{v}_2) \xrightarrow{d} 2U - 2V,$$

where U and V are independent (due to the sample splitting) with $U \sim \chi_2^2$ and $V \sim \chi_1^2$. It turns out that the denominator of T_{W} can also affect the limiting behaviour of T_{W} non-trivially in this degenerate situation, and the exact form of the limiting distribution relies on the choice of $\widehat{\eta}_1$ and $\widehat{\eta}_2$. Nonetheless, we can still argue that the limiting distribution is not Gaussian. To see this, note that

$$\mathbb{P}_P(T_{\mathrm{W}} < 0) \to \mathbb{P}_P\Big(U - V < 0\Big) = \mathbb{P}_P\Big(F_{2,1} < 1/2\Big) \neq \Phi(0).$$

Therefore, when $\eta_1 = \eta_2 = 0$, the test based on T_{W} can be either conservative or anti-conservative depending on the choice of α.

3.10 Splines

This section is a self-contained description of spline spaces and some of their properties relevant for the spline regressions in Section 3.5. The definitions given here are not standard in the spline literature, in the sense that they are less general than the usual definitions, but they suffice for the purposes of regression with splines. One particular simplification that we will adhere to throughout is to restrict attention to splines with equispaced knots that are defined on the unit hypercube.

We start by considering function spaces of piecewise polynomials with adjustable degrees of smoothness and give a definition of uniform B-splines.

Definition 3.1. Let $N \in \mathbb{N}_0$, and let $\Delta = (\Delta_\ell)_{\ell=0}^{N+1}$ be the knots of an equispaced partition of $[0,1]$, with $\Delta_0 := 0$ and $\Delta_{N+1} := 1$. For $r \in \mathbb{N}$, define the *spline space* $\mathcal{S}_{r,N}$ to be the set of functions $f : [0,1] \to \mathbb{R}$, where the restriction of f to $[\Delta_{\ell-1}, \Delta_\ell]$ is a polynomial of degree at most $r-1$ for $\ell \in [N+1]$ and where f is $(r-2)$-times continuously differentiable when $r \geq 2$ (we interpret this as meaning 'continuous' when $r = 2$). We say that r is the *order* of $\mathcal{S}_{r,N}$. Define the vector $t = (t_1, \dots, t_{N+2r}) \in [0,1]^{N+2r}$ by

$$t := (\underbrace{\Delta_0, \dots, \Delta_0}_{r}, \Delta_1, \dots, \Delta_N \underbrace{\Delta_{N+1}, \dots, \Delta_{N+1}}_{r}).$$

For $s \in [r]$ and $k \in [N + 2r - s]$, define the functions $B_{k,s,N}$ recursively for $x \in [0,1)$ by

$$B_{k,1,N}(x) := \mathbb{1}_{[t_k, t_{k+1})}(x),$$

and, for $s \in \{2, \dots, r\}$,

$$B_{k,s,N}(x) := \frac{x - t_k}{t_{k+s-1} - t_k} B_{k,s-1,N}(x) + \frac{t_{k+s} - x}{t_{k+s} - t_{k+1}} B_{k+1,s-1,N}(x)$$

(with the convention that $0/0 := 0$). We also define $B_{k,s,N}(1) := \lim_{x \nearrow 1} B_{k,s,N}(x)$. The $K := N + r$ functions $B_{1,r,N}, \dots, B_{K,r,N}$ are called *B-splines*.

It is standard in the spline literature to parametrise spline spaces in terms of the order r of the polynomials rather than the degree $(r-1)$. The Curry–Schoenberg theorem gives a relationship between the two definitions above.

Proposition 3.7 (Curry–Schoenberg)**.** *The set of B-splines $\{B_{k,r,N}\}_{k=1}^K$ is a basis for $\mathcal{S}_{r,N}$.*

Proof. See Schumaker (2007, Theorem 4.13). □

It is worth noting some properties of B-splines and the spline space $\mathcal{S}_{r,N}$.

Proposition 3.8. *(a) The B-splines $\{B_{k,r,N}\}_{k=1}^K$ are non-negative and form a partition of unity; i.e.,*

$$\sum_{k=1}^{K} B_{k,r,N}(x) = 1,$$

for all $x \in [0,1]$.

(b) For any $f \in \mathcal{S}_{r,N}$ *of the form* $f(x) = \sum_{k=1}^{K} \beta_k B_{k,r,N}(x)$ *with* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K) \in \mathbb{R}^K$*, there exists* $c_s(r) > 0$*, depending only on* r*, such that*

$$c_s(r) K^{-1/p} \|\boldsymbol{\beta}\|_p \leq \|f\|_p \leq 2^{1/p} K^{-1/p} \|\boldsymbol{\beta}\|_p, \tag{3.54}$$

for all $p \in [1, \infty]$*. In particular,*

$$c_s(r) K^{-1/p} \leq \|B_{k,r,N}\|_p \leq 2^{1/p} K^{-1/p}$$

for all $k \in [K]$.

Proof. (a) This follows from Equations (4.5) and (4.10) of de Boor (1976).

(b) The conclusion of de Boor (1976, Theorem 5.2) yields the existence of $c_s(r) > 0$ such that

$$c_s(r)(N+1)^{-1/p} \|\boldsymbol{\beta}\|_p \leq \|f\|_p \leq r^{-1/p}(N+1)^{-1/p} \|\boldsymbol{\beta}\|_p. \tag{3.55}$$

But $(N+1)^{-1/p} \geq K^{-1/p}$ since $K = N + r$, yielding the lower bound in (3.54). For the upper bound in (3.54), we note that

$$K \leq 2 \max(N, r) \leq 2(N+1)r \tag{3.56}$$

and rearranging yields the desired result. □

We will require splines on $[0,1]^d$ instead of just $[0,1]$, and to that end we tensorise our earlier spline constructions.

Definition 3.2. Recall Definition 3.1. Let $d \in \mathbb{N}$ and define the *d-tensor spline space*

$$\mathcal{S}_{r,N}^d := \left\{ f : [0,1]^d \to \mathbb{R}, f(x_1, \ldots, x_d) = \prod_{j=1}^{d} f_j(x_j) : f_j \in \mathcal{S}_{r,N} \ \forall j \in [d] \right\}.$$

Let $\otimes$ denote the vectorised outer product operator, so that $x \otimes y := \mathrm{vec}(xy^\top)$ for Euclidean vectors x and y, where vec denotes the vectorisation operator. Define $\boldsymbol{B}_{r,N} : [0,1] \to \mathbb{R}^{N+r}$ with kth component function $B_{k,r,N}$, so that $\boldsymbol{B}_{r,N}(x) = (B_{1,r,N}(x), \ldots, B_{N+r,r,N}(x))^\top$. Now redefine $K := (N+r)^d$; since $\otimes$ is associative, we may define tensor-basis functions $\boldsymbol{\phi} \equiv \boldsymbol{\phi}_{r,N} : [0,1]^d \to \mathbb{R}^K$ by

$$\boldsymbol{\phi}(x_1, \ldots, x_d) \equiv (\phi_1(x_1, \ldots, x_d), \ldots, \phi_K(x_1, \ldots, x_d))^\top := \boldsymbol{B}_{r,N}(x_1) \otimes \cdots \otimes \boldsymbol{B}_{r,N}(x_d).$$

By properties of the tensor product and the Curry–Schoenberg theorem, the collection $\{\phi_k\}_{k=1}^K$ forms a basis for the d-tensor spline space $\mathcal{S}_{r,N}^d$ under the usual pointwise addition and scalar multiplication operations, and we refer to it as the *d-tensor B-spline basis of* $\mathcal{S}_{r,N}^d$. In our subsequent asymptotic results, the first of which is Lemma 3.13, we will think of d and r

as fixed, but allow N (and consequently K) to depend on n. Proposition 3.9 below shows that the properties of univariate B-splines given in Proposition 3.8 carry over to the d-tensor splines.

Proposition 3.9. *(a) The basis functions $\{\phi_k\}_{k=1}^K$ are non-negative and form a partition of unity.*

(b) For any $f \in \mathcal{S}^d_{r,N}$ of the form $f(x) = \sum_{k=1}^K \beta_k \phi_k(x)$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K) \in \mathbb{R}^K$, and for any $p \in [1, \infty]$,

$$c_s(r)^d K^{-1/p} \|\boldsymbol{\beta}\|_p \leq \|f\|_p \leq 2^{d/p} K^{-1/p} \|\boldsymbol{\beta}\|_p,$$

where $c_s(r) > 0$ is taken from Proposition 3.8(b). In particular, for all $k \in [K]$,

$$c_s(r)^d K^{-1/p} \leq \|\phi_k\|_p \leq 2^{d/p} K^{-1/p}.$$

(c) For any Z with distribution P on $[0,1]^d$, the matrix $\boldsymbol{\Sigma}_P := \mathbb{E}_P(\boldsymbol{\phi}(Z)\boldsymbol{\phi}(Z)^\top) \in \mathbb{R}^{K \times K}$ satisfies

$$\lambda_{\min}(\boldsymbol{\Sigma}_P) \leq K^{-1}$$

and

$$\lambda_{\max}(\boldsymbol{\Sigma}_P) \geq K^{-2}.$$

(d) Now suppose that P is absolutely continuous with respect to Lebesgue measure on $[0,1]^d$ with density p. If $C := \sup_{z \in [0,1]^d} p(z) < \infty$, then

$$\lambda_{\max}(\boldsymbol{\Sigma}_P) \leq C 2^d K^{-1}.$$

If instead $c := \inf_{z \in [0,1]^d} p(z) > 0$, then

$$\lambda_{\min}(\boldsymbol{\Sigma}_P) \geq c c_s(r)^d K^{-1},$$

where $c_s(r) > 0$ is taken from (b).

Proof. *(a)* This follows from Proposition 3.8*(a)* and the definition of the d-tensor B-spline basis.

(b) We will only prove the case $d = 2$, since the full result will then follow by induction on d. To prove the lower bound, we can write

$$f(x_1, x_2) = \sum_{k=1}^{\sqrt{K}} \sum_{\ell=1}^{\sqrt{K}} \beta_{k\ell} B_{k,r,N}(x_1) B_{\ell,r,N}(x_2) =: \sum_{k=1}^{\sqrt{K}} \gamma_k(x_2) B_{k,r,N}(x_1).$$

For $p \in [1, \infty)$, we have by using (3.55) twice that

$$\|f\|_p^p = \int_0^1 \int_0^1 \Big| \sum_{k=1}^{\sqrt{K}} \gamma_k(x_2) B_{k,r,N}(x_1) \Big|^p \, \mathrm{d}x_1 \, \mathrm{d}x_2 \geq \frac{c_s(r)^p}{N+1} \int_0^1 \sum_{k=1}^{\sqrt{K}} |\gamma_k(x_2)|^p \, \mathrm{d}x_2$$

$$\geq \frac{c_s(r)^{2p}}{(N+1)^2} \sum_{k=1}^{\sqrt{K}} \sum_{\ell=1}^{\sqrt{K}} |\beta_{k\ell}|^p = \frac{c_s(r)^{2p}}{(N+1)^2} \|\boldsymbol{\beta}\|_p^p \geq \frac{c_s(r)^{2p}}{K^2} \|\boldsymbol{\beta}\|_p^p,$$

as desired. For $p = \infty$, we have by a similar argument that

$$\begin{aligned}\|f\|_\infty &= \sup_{x_1,x_2\in[0,1]}\left|\sum_{k=1}^{\sqrt{K}}\gamma_k(x_2)B_{k,r,N}(x_1)\right| \geq c_s(r)\sup_{x_2\in[0,1]}\max_{k\in[\sqrt{K}]}|\gamma_k(x_2)| \\ &\geq c_s(r)^2\max_{k,\ell\in[\sqrt{K}]}|\beta_{k\ell}| = c_s(r)^2\|\boldsymbol{\beta}\|_\infty,\end{aligned}$$

again as desired. To prove the upper bound, we argue very similarly, and use the fact that, with K redefined as $(N+r)^d$, we have $K \leq 2^d(N+1)^d r^d$ by (3.56).

(c) Note that

$$K\lambda_{\min}(\boldsymbol{\Sigma}_P) \leq \operatorname{tr}(\boldsymbol{\Sigma}_P) = \mathbb{E}\Big(\sum_{k=1}^K \phi_k^2(Z)\Big) \leq \mathbb{E}\bigg(\Big\{\sum_{k=1}^K \phi_k(Z)\Big\}^2\bigg) = 1,$$

and by Cauchy–Schwarz,

$$K\lambda_{\max}(\boldsymbol{\Sigma}_P) \geq \operatorname{tr}(\boldsymbol{\Sigma}_P) = \mathbb{E}\Big(\sum_{k=1}^K \phi_k^2(Z)\Big) \geq \frac{1}{K}\mathbb{E}\bigg(\Big\{\sum_{k=1}^K \phi_k(Z)\Big\}^2\bigg) = \frac{1}{K}.$$

(d) We have

$$\begin{aligned}\lambda_{\max}(\boldsymbol{\Sigma}_P) &= \sup_{\boldsymbol{\beta}\in\mathbb{R}^K:\|\boldsymbol{\beta}\|_2=1}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_P\boldsymbol{\beta} = \sup_{\boldsymbol{\beta}\in\mathbb{R}^K:\|\boldsymbol{\beta}\|_2=1}\mathbb{E}\bigg(\Big\{\sum_{k=1}^K\beta_k\phi_k(Z)\Big\}^2\bigg) \\ &\leq C\sup_{\boldsymbol{\beta}\in\mathbb{R}^K:\|\boldsymbol{\beta}\|_2=1}\Big\|\sum_{k=1}^K\beta_k\phi_k\Big\|_2^2 \leq \frac{2^dC}{K},\end{aligned}$$

where the final inequality uses *(b)*. By a similar argument,

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\Sigma}_P) &= \inf_{\boldsymbol{\beta}\in\mathbb{R}^K:\|\boldsymbol{\beta}\|_2=1}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_P\boldsymbol{\beta} = \inf_{\boldsymbol{\beta}\in\mathbb{R}^K:\|\boldsymbol{\beta}\|_2=1}\mathbb{E}\bigg(\Big\{\sum_{k=1}^K\beta_k\phi_k(Z)\Big\}^2\bigg) \\ &\geq c\inf_{\boldsymbol{\beta}\in\mathbb{R}^K:\|\boldsymbol{\beta}\|_2=1}\Big\|\sum_{k=1}^K\beta_k\phi_k\Big\|_2^2 \geq \frac{c}{c_s(r)^dK},\end{aligned}$$

as required. □

We will now argue that splines are strong approximators over classes of sufficiently smooth functions, as defined by Hölder smoothness:

Definition 3.3. Given a multi-index $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_d) \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| := \sum_{j=1}^d \alpha_j$ and an $|\boldsymbol{\alpha}|$-times differentiable function $f : [0,1]^d \to \mathbb{R}$, we define

$$D^{\boldsymbol{\alpha}}f := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}}\cdots\frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}f.$$

For $s > 0$, write $s_0 := \lceil s \rceil - 1$ and define $\mathcal{H}_s \equiv \mathcal{H}_s^d$ to be the set $f : [0,1]^d \to \mathbb{R}$ that are s_0-times differentiable and that satisfy

$$\max_{\boldsymbol{\alpha} \in \mathbb{N}_0^d : |\boldsymbol{\alpha}| = s_0} |D^{\boldsymbol{\alpha}} f(x) - D^{\boldsymbol{\alpha}} f(\tilde{x})| \leq C \|x - \tilde{x}\|_2^{s - s_0} \quad \forall x, \tilde{x} \in [0,1]^d \tag{3.57}$$

and

$$\max_{\boldsymbol{\alpha} \in \mathbb{N}_0^d : |\boldsymbol{\alpha}| = s_0} \|D^{\boldsymbol{\alpha}} f\|_\infty \leq C \tag{3.58}$$

for some $C > 0$. If $f \in \mathcal{H}_s$, then the infimum of the set of $C > 0$ for which both (3.57) and (3.58) hold is called the s-Hölder norm, and is denoted by $\|f\|_{\mathcal{H}_s}$.

The following basic result shows that given normed space of real-valued functions on $[0,1]^d$ containing $\mathcal{S}_{r,N}^d$, we can find a best approximant within $\mathcal{S}_{r,N}^d$.

Lemma 3.12. *Let $(\mathcal{V}, \|\cdot\|)$ denote a normed space of real-valued functions on $[0,1]^d$ that contains $\mathcal{S}_{r,N}^d$ as a subspace. Then given any $f \in \mathcal{V}$, there exists $f^* \in \mathcal{S}_{r,N}^d$ such that $\|f - f^*\| = \inf_{g \in \mathcal{S}_{r,N}^d} \|f - g\|$. If $\|\cdot\|$ is strictly convex, then this best approximant is unique.*

Proof. Since $\mathcal{S}_{r,N}^d$ is a finite-dimensional subspace of $\mathcal{V}$, Powell (1981, Theorem 1.2) guarantees the existence of the best approximant f^*. The uniqueness property follows from Powell (1981, Theorem 2.4) since $\mathcal{S}_{r,N}^d$ is convex. □

The approximation properties of splines over Hölder smoothness classes are characterised below.

Proposition 3.10. *Suppose $f \in \mathcal{H}_s^d$. Then there exists $C(d,r) > 0$ and $f^* \in \mathcal{S}_{r,N}^d$ such that*

$$\|f - f^*\|_\infty \leq \frac{C(d,r)}{(N+1)^{\min(s,r)}} \|f\|_{\mathcal{H}_s} \leq \frac{C(d,r)}{(2rK)^{\min(s,r)/d}} \|f\|_{\mathcal{H}_s}. \tag{3.59}$$

Proof. Given $g : [0,1]^d \to \mathbb{R}$, $j \in [d]$, $h > 0$ and $k \in \mathbb{N}$, we define the *kth forward difference of g in coordinate j with spacing h at x* by

$$\Delta_{j,h}^k g(x) := \sum_{\ell=0}^{k} (-1)^{k-\ell} \binom{k}{\ell} g(x + \ell h e_j),$$

where $e_j \in \mathbb{R}^d$ denotes the jth standard basis vector. The *kth modulus of smoothness of g in coordinate j of radius $t \in (0, 1/k]$* is then defined as

$$\omega_j^k(g; t) := \sup_{h \in [0,t]} \sup_{x \in [0, 1-kh]} |\Delta_{j,h}^k g(x)|.$$

By Lemma 3.12 and Schumaker (2007, Theorem 12.8), there exists $f^* \in \mathcal{S}_{r,N}^d$ such that

$$\|f - f^*\|_\infty \leq C'(d,r) \sum_{j=1}^{d} \omega_j^r(f; 1/(N+1)),$$

for some $C'(d,r) > 0$ depending only on d and r. First consider the case $r \geq s$, and recall the notation $s_0 := \lceil s \rceil - 1$. By Schumaker (2007, (2.119) and (2.117) in Theorem 2.59),

$$\begin{aligned}\omega_j^r(f; 1/(N+1)) &\leq \frac{1}{(N+1)^{s_0}} \omega_j^{r-s_0}(D^{s_0 e_j} f; 1/(N+1)) \\ &\leq \frac{2^{r-s_0-1}}{(N+1)^{s_0}} \omega_j^1(D^{s_0 e_j} f; 1/(N+1)) \leq \frac{2^{r-1}}{(N+1)^s} \|f\|_{\mathcal{H}_s}.\end{aligned}$$

On the other hand, if $r < s$, then by Schumaker (2007, (2.120) in Theorem 2.59),

$$\omega_j^r(f; 1/(N+1)) \leq \frac{1}{(N+1)^r} \|D^{r e_j} f\|_\infty \leq \frac{1}{(N+1)^r} \|f\|_{\mathcal{H}_s}.$$

Combining these bounds yields the first inequality in (3.59) with $C(d,r) := 2^{r-1} d C'(d,r)$. The final bound again follows from the fact that $K \leq 2^d (N+1)^d r^d$. □

Our next result provides a way of translating properties between population least squares approximants and supremum norm approximants.

Proposition 3.11. *Let Z be a random vector taking values in $[0,1]^d$, and let $\mathcal{F}$ be a class of functions $f : [0,1]^d \to \mathbb{R}$ with $\mathbb{E}(f(Z)^2) < \infty$ for all $f \in \mathcal{F}$. Then for each $f \in \mathcal{F}$, there exists a unique $f^\dagger \in \mathcal{S}_{r,N}^d$ such that*

$$\mathbb{E}(\{f(Z) - f^\dagger(Z)\}^2) = \inf_{g \in \mathcal{S}_{r,N}^d} \mathbb{E}(\{f(Z) - g(Z)\}^2).$$

Now fix $f \in \mathcal{F}$ and $f^* \in \mathcal{S}_{r,N}^d$. Further, suppose that Z has a density p with respect to Lebesgue measure on $[0,1]^d$ satisfying $c := \inf_{z \in [0,1]^d} p(z) > 0$ and $C := \sup_{z \in [0,1]^d} p(z) < \infty$. Then there exists $M(c, C, d, r) > 0$ such that

$$\|f - f^\dagger\|_\infty \leq M(c, C, d, r) \|f - f^*\|_\infty.$$

Proof. Let P denote the distribution of Z, and let $L_2(P)$ denote the normed space of equivalence classes of measurable functions $g : [0,1]^d \to \mathbb{R}$ satisfying

$$\|g\|_{2,P} := \{\mathbb{E}(g(Z)^2)\}^{1/2} < \infty$$

under the binary relation where $g \sim g^\circ$ if $g(Z) = g^\circ(Z)$ almost surely[1]. The existence of the unique $f^\dagger \in \mathcal{S}_{r,N}^d$ follows from Lemma 3.12 since the $L_2(P)$ norm is strictly convex.

Now define $\tilde{g} := f - f^*$, so the unique $L_2(P)$-best approximant $\tilde{g}^\dagger$ to $\tilde{g}$ in $\mathcal{S}_{r,N}^d$ is given by $\tilde{g}^\dagger = f^\dagger - f^*$. We now verify that Conditions A.1, A.2 and A.3 of Huang (2003, Theorem A.1) hold. Condition A.1 is satisfied by our hypotheses on c, C; Condition A.2 holds since the knots of the splines in $\mathcal{S}_{r,N}^d$ are equispaced; and Condition A.3 is satisfied by Proposition 3.9(b), where we again use the bounds $(N+1)^d \leq K \leq 2^d (N+1)^d r^d$. Thus, Huang (2003, Theorem

[1] We do not distinguish between a function with finite $\|\cdot\|_{2,P}$ norm and its equivalence class in what follows.

A.1) yields the existence of $M'(c, C, d, r) > 0$ such that

$$\|\tilde{g}^\dagger\|_\infty \leq M'(c, C, d, r)\|\tilde{g}\|_\infty.$$

We conclude that

$$\|f - f^\dagger\|_\infty \leq \|\tilde{g}\|_\infty + \|\tilde{g}^\dagger\|_\infty \leq (1 + M'(c, C, d, r))\|\tilde{g}\|_\infty,$$

so the desired result holds with $M(c, C, d, r) := 1 + M'(c, C, d, r)$. □

Lemma 3.13 below will ensure that, provided K increases slightly slower than n (so that $K\log(K)/n \to 0$), performing ordinary least squares with the d-tensor B-spline basis will yield consistent estimators.

Lemma 3.13. *Let $\mathcal{P}$ denote a family of distributions for a random vector Z taking values in $[0,1]^d$, and let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed copies of X. Recall the notation $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)^\top$, where $\{\phi_k\}_{k=1}^K$ denotes the d-tensor B-spline basis of $\mathcal{S}^d_{r,N}$. For $P \in \mathcal{P}$, define $\boldsymbol{\Sigma}_P := \mathbb{E}_P(\boldsymbol{\phi}(Z)\boldsymbol{\phi}(Z)^\top) \in \mathbb{R}^{K\times K}$ and $\widehat{\boldsymbol{\Sigma}} := n^{-1}\sum_{i=1}^n \boldsymbol{\phi}(Z_i)\boldsymbol{\phi}(Z_i)^\top$. Suppose that each $P \in \mathcal{P}$ is absolutely continuous with respect to Lebesgue measure on $[0,1]^d$, with corresponding density p_P satisfying $C := \sup_{P\in\mathcal{P}} \sup_{z\in[0,1]^d} p_P(z) < \infty$. Then*

$$K\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{op} = O_{\mathcal{P}}\Big(\frac{K\log(eK)}{n} + \sqrt{\frac{K\log(eK)}{n}}\Big).$$

If, in addition, $c := \inf_{P\in\mathcal{P}} \inf_{z\in[0,1]^d} p_P(z) > 0$ and $K\log(K)/n \to 0$, then

$$K^{-1}\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}_P^{-1}\|_{op} = O_{\mathcal{P}}\Big(\frac{K\log(eK)}{n} + \sqrt{\frac{K\log(eK)}{n}}\Big),$$

and

$$\|\widehat{\boldsymbol{\Sigma}}\|_{op} = O_{\mathcal{P}}(K^{-1}), \quad \|\widehat{\boldsymbol{\Sigma}}^{-1}\|_{op} = O_{\mathcal{P}}(K). \tag{3.60}$$

Proof. For the first claim, by Markov's inequality, it suffices to show that

$$\sup_{P\in\mathcal{P}} K\mathbb{E}_P(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}}) = O\Big(\frac{K\log(eK)}{n} + \sqrt{\frac{K\log(eK)}{n}}\Big)$$

as $n \to \infty$. By the Rudelson law of large numbers for matrices (Belloni et al., 2015, Lemma 6.2) (and Chebyshev's inequality when $K = 1$), there exists a universal constant $C_* > 0$ such that

$$\begin{aligned}\sup_{P\in\mathcal{P}} K\mathbb{E}_P(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}}) &\leq \frac{C_*K\log(eK)}{n} + C_*\sqrt{\frac{K^2\log(eK)}{n}} \sup_{P\in\mathcal{P}} \sqrt{\|\boldsymbol{\Sigma}_P\|_{\mathrm{op}}} \\ &\leq \frac{C_*K\log(eK)}{n} + C_*\sqrt{C2^d}\sqrt{\frac{K\log(eK)}{n}},\end{aligned}$$

since $\|\boldsymbol{\phi}(Z)\|_2 \le \|\boldsymbol{\phi}(Z)\|_1 = 1$ and $\|\boldsymbol{\Sigma}_P\|_{\mathrm{op}} \le C2^d K^{-1}$, by Proposition 3.9(a) and (d) respectively.

For the second claim, note first that $K\lambda_{\min}(\boldsymbol{\Sigma}_P) \ge cc_s(r) =: b > 0$ by Proposition 3.9(d), and by Weyl's inequality, $K\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}) \ge b - K\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}}$. Thus, on the event $\{K\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}} \le b/2\}$, we have that $\widehat{\boldsymbol{\Sigma}}$ is invertible, and

$$\begin{aligned}\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}} &= \|\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\Sigma}_P - \widehat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}} \\ &\le (\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}} + \|\boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}})\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}}\|\boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}}.\end{aligned}$$

Since, on the event $\{K\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}} \le b/2\}$, we have $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}}\|\boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}} \le 1/2$, we deduce that on this event,

$$\begin{aligned}\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}} &\le \frac{\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}}\|\boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}}^2}{1 - \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}}\|\boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}}} \\ &\le 2\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}}\|\boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}}^2 \le \frac{2K^2}{b^2}\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}}.\end{aligned}$$

From the first claim of the lemma and the hypothesis that $K\log(K)/n \to 0$, given $\epsilon > 0$, we can choose $n_0 \in \mathbb{N}$ large enough that

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P\left(K\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}} \ge \frac{b}{2}\right) \le \frac{\epsilon}{2}$$

for $n \ge n_0$. Then, by another application of the first claim of the lemma, by increasing n_0 if necessary, we can find $M_0 > 0$ such that

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P\left(K\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}} \ge \frac{b^2 M}{2}\left\{\frac{K\log(eK)}{n} + \sqrt{\frac{K\log(eK)}{n}}\right\}\right) \le \frac{\epsilon}{2}$$

for all $n \ge n_0$ and $M \ge M_0$. It follows that for $n \ge n_0$ and $M \ge M_0$, we have

$$\begin{aligned}&\sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\frac{1}{K}\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}} \ge M\left\{\frac{K\log(eK)}{n} + \sqrt{\frac{K\log(eK)}{n}}\right\}\right) \\ &\quad\le \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(K\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}} \ge \frac{b^2 M}{2}\left\{\frac{K\log(eK)}{n} + \sqrt{\frac{K\log(eK)}{n}}\right\}\right) \\ &\qquad + \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(K\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}} > \frac{b}{2}\right) \le \epsilon,\end{aligned}$$

which establishes the second claim.

Finally, by the first part of Proposition 3.9(d),

$$K\|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}} \le K\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_P\|_{\mathrm{op}} + K\|\boldsymbol{\Sigma}_P\|_{\mathrm{op}} = O_{\mathcal{P}}(1),$$

and by the second part of Proposition 3.9(d),

$$K^{-1}\|\widehat{\boldsymbol{\Sigma}}^{-1}\|_{\mathrm{op}} \leq K^{-1}\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}} + K^{-1}\|\boldsymbol{\Sigma}_P^{-1}\|_{\mathrm{op}} = O_{\mathcal{P}}(1),$$

as required. □

Proposition 3.12 below provides estimation and both in-sample and out-of-sample prediction bounds for spline regression. It is based on Belloni et al. (2015, Theorem 4.1), but here we control the errors in a uniform fashion over a family of distributions, and those authors did not require in-sample bounds.

Proposition 3.12. *Let $\mathcal{P}$ be a family of distributions of (Y, Z) on $\mathbb{R} \times [0,1]^d$ with regression function f_P given by $f_P(z) := \mathbb{E}_P(Y \mid Z = z)$, and let $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ be independent and identically distributed copies of (Y, Z). Suppose that*

(i) The $L_2(P)$-best approximant $f_P^\dagger$ of f_P in $\mathcal{S}_{r,N}^d$ satisfies

$$\sup_{P\in\mathcal{P}} \|f_P - f_P^\dagger\|_\infty = O(K^{-\zeta}),$$

for some $\zeta \equiv \zeta(d, r) > 0$.

(ii) Each $P \in \mathcal{P}$ is absolutely continuous with respect to Lebesgue measure on $[0,1]^d$, with corresponding density p_P satisfying

$$C := \sup_{P\in\mathcal{P}} \sup_{z\in[0,1]^d} p_P(z) < \infty \quad \text{and} \quad c := \inf_{P\in\mathcal{P}} \inf_{z\in[0,1]^d} p_P(z) > 0.$$

(iii) There exists a positive sequence $(\sigma_n^2)_{n\in\mathbb{N}}$ such that $\mathrm{Var}_P(Y \mid Z) \leq \sigma_n^2$ for all $P \in \mathcal{P}$.

Let $\boldsymbol{\phi}$ denote the d-tensor B-spline basis of $\mathcal{S}_{r,N}^d$ and let $\widehat{\boldsymbol{\beta}}$ denote the ordinary least squares estimate from regressing $Y_1, \ldots, Y_n$ onto $\boldsymbol{\phi}(Z_1), \ldots, \boldsymbol{\phi}(Z_n)$. Assume that $K\log(K)/n \to 0$. Then

$$\frac{1}{n}\sum_{i=1}^n \big(f_P(Z_i) - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\phi}(Z_i)\big)^2 = O_{\mathcal{P}}\big(K^{-2\zeta} + \sigma_n^2 K/n\big).$$

Letting $\boldsymbol{\beta}_P \in \mathbb{R}^K$ be the unique solution to $f_P^\dagger(z) = \boldsymbol{\beta}_P^\top \boldsymbol{\phi}(z)$, we have under the same assumptions that

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_P\|_2^2 = O_{\mathcal{P}}\big(K^{-(2\zeta-2)}/n + \sigma_n^2 K^2/n\big).$$

Further, if (Y^*, Z^*) is a new observation of (Y, Z) independent of the original sample, then

$$\mathbb{E}_P\Big(\big\{f_P(Z^*) - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\phi}(Z^*)\big\}^2 \mid \widehat{\boldsymbol{\beta}}\Big) = O_{\mathcal{P}}\big(K^{-2\zeta} + \sigma_n^2 K/n\big).$$

Proof. Let $\widehat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}(Z_i)\boldsymbol{\phi}(Z_i)^\top$ and for $i \in [n]$, let $h_i := f_P(Z_i) - f_P^\dagger(Z_i)$ and $\varepsilon_i := Y_i - f_P(Z_i)$. Then, recalling that $f_P^\dagger(z) = \boldsymbol{\beta}_P^\top \boldsymbol{\phi}(z)$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \big(f_P(Z_i) - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\phi}(Z_i)\big)^2 \le 2\|f_P - f_P^\dagger\|_\infty^2 + 2(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_P)^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_P) \\
&= 2\|f_P - f_P^\dagger\|_\infty^2 + 2\bigg(\frac{1}{n}\sum_{i=1}^n (h_i + \varepsilon_i)\boldsymbol{\phi}(Z_i)\bigg)^\top \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}^{-1}\bigg(\frac{1}{n}\sum_{i=1}^n (h_i + \varepsilon_i)\boldsymbol{\phi}(Z_i)\bigg) \\
&\le 2\|f_P - f_P^\dagger\|_\infty^2 + 2\|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}\big\|\widehat{\boldsymbol{\Sigma}}^{-1}\big\|_{\mathrm{op}}\bigg\|\widehat{\boldsymbol{\Sigma}}^{-1/2}\bigg(\frac{1}{n}\sum_{i=1}^n (h_i + \varepsilon_i)\boldsymbol{\phi}(Z_i)\bigg)\bigg\|_2^2. \qquad (3.61)
\end{aligned}$$

Now $\|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}\big\|\widehat{\boldsymbol{\Sigma}}^{-1}\big\|_{\mathrm{op}} = O_{\mathcal{P}}(1)$ by (3.60) in Lemma 3.13. Moreover,

$$\begin{aligned}
\bigg\|\widehat{\boldsymbol{\Sigma}}^{-1/2}\bigg(\frac{1}{n}\sum_{i=1}^n (h_i + \varepsilon_i)\boldsymbol{\phi}(Z_i)\bigg)\bigg\|_2^2 \le 2\underbrace{\bigg\|\widehat{\boldsymbol{\Sigma}}^{-1/2}\bigg(\frac{1}{n}\sum_{i=1}^n h_i\boldsymbol{\phi}(Z_i)\bigg)\bigg\|_2^2}_{\mathrm{I}_n} \\
+ 2\underbrace{\bigg\|\widehat{\boldsymbol{\Sigma}}^{-1/2}\bigg(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\boldsymbol{\phi}(Z_i)\bigg)\bigg\|_2^2}_{\mathrm{II}_n}.
\end{aligned}$$

To deal with I_n, let $\boldsymbol{\Sigma}_P := \mathbb{E}_P(\boldsymbol{\phi}(Z)\boldsymbol{\phi}(Z)^\top) \in \mathbb{R}^{K\times K}$, and note that

$$\begin{aligned}
\mathbb{E}_P(f_P^\dagger(Z)\boldsymbol{\phi}(Z)^\top) &= \mathbb{E}_P(\boldsymbol{\beta}_P^\top\boldsymbol{\phi}(Z)\boldsymbol{\phi}(Z)^\top) \\
&= \mathbb{E}_P(f_P(Z)\boldsymbol{\phi}(Z)^\top)\boldsymbol{\Sigma}_P^{-1}\mathbb{E}_P(\boldsymbol{\phi}(Z)\boldsymbol{\phi}(Z)^\top) = \mathbb{E}_P(f_P(Z)\boldsymbol{\phi}(Z)^\top). \qquad (3.62)
\end{aligned}$$

It follows that

$$\begin{aligned}
\sup_{P\in\mathcal{P}} \mathbb{E}_P\bigg(\bigg\|\frac{1}{n}\sum_{i=1}^n h_i\boldsymbol{\phi}(Z_i)\bigg\|_2^2\bigg) &= \frac{1}{n}\sup_{P\in\mathcal{P}} \mathbb{E}_P\{\mathrm{tr}((h_1)^2\boldsymbol{\phi}(Z_1)\boldsymbol{\phi}(Z_1)^\top)\} \\
&\le \frac{1}{n}\sup_{P\in\mathcal{P}} \|f_P - f_P^\dagger\|_\infty^2 \mathrm{tr}(\boldsymbol{\Sigma}_P), \qquad (3.63)
\end{aligned}$$

so

$$|\mathrm{I}_n| \le \big\|\widehat{\boldsymbol{\Sigma}}^{-1}\big\|_{\mathrm{op}}\bigg\|\frac{1}{n}\sum_{i=1}^n h_i\boldsymbol{\phi}(Z_i)\bigg\|_2^2 = O_{\mathcal{P}}\big(K^{-(2\zeta-1)}/n\big) \qquad (3.64)$$

by our assumption on $\sup_{P\in\mathcal{P}}\|f_P - f_P^\dagger\|_\infty$, Proposition 3.9(d), (3.60) in Lemma 3.13 and Lemma 3.2. To deal with II_n, we note that $\varepsilon_1, \ldots, \varepsilon_n$ are conditionally independent given $Z_1, \ldots, Z_n$, so

$$\mathbb{E}_P(\mathrm{II}_n \,|\, Z_1, \ldots, Z_n) = \frac{1}{n^2}\mathrm{tr}\bigg(\widehat{\boldsymbol{\Sigma}}^{-1}\sum_{i=1}^n \mathbb{E}(\varepsilon_i^2 \,|\, Z_i)\boldsymbol{\phi}(Z_i)\boldsymbol{\phi}(Z_i)^\top\bigg) \le \frac{\sigma_n^2}{n}\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Sigma}}) \le \frac{\sigma_n^2 K}{n}. \qquad (3.65)$$

Putting things together, since $\|f_P - f_P^\dagger\|_\infty^2 = O(K^{-2\zeta})$ by assumption, we have

$$\frac{1}{n}\sum_{i=1}^{n}(f_P(Z_i) - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\phi}(Z_i))^2 = O_{\mathcal{P}}(K^{-2\zeta} + \sigma_n^2 K/n),$$

as desired.

For the second claim, observe that

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_P\|_2^2 &= \left\|\widehat{\boldsymbol{\Sigma}}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(h_i + \varepsilon_i)\boldsymbol{\phi}(Z_i)\right)\right\|_2^2 \\ &\leq \|\widehat{\boldsymbol{\Sigma}}^{-1}\|_{\mathrm{op}}\left\|\widehat{\boldsymbol{\Sigma}}^{-1/2}\left(\frac{1}{n}\sum_{i=1}^{n}(h_i + \varepsilon_i)\boldsymbol{\phi}(Z_i)\right)\right\|_2^2 = O_{\mathcal{P}}(K^{-(2\zeta-2)}/n + \sigma_n^2 K^2/n),\end{aligned}$$

by (3.60) in Lemma 3.13 and our results above.

Finally, we have following the argument in (3.61), we have

$$\begin{aligned}\mathbb{E}_P\left(\{f_P(Z^*) - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\phi}(Z^*)\}^2 \mid \widehat{\boldsymbol{\beta}}\right) &\leq 2\|f_P - f_P^\dagger\|_\infty^2 + 2\|\boldsymbol{\Sigma}_P\|_{\mathrm{op}}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_P\|^2 \\ &= O_{\mathcal{P}}(K^{-2\zeta} + \sigma_n^2 K/n),\end{aligned}$$

by Proposition 3.9(d) and the second claim of the proposition. □

Under standard smoothness assumptions, we can derive the following consequence of Proposition 3.12:

Corollary 3.2. *Let $\mathcal{P}$ be a family of distributions of (Y, Z) on $\mathbb{R} \times [0,1]^d$, and let f_P denote the regression function given by $f_P(z) := \mathbb{E}_P(Y \mid Z = z)$ for $P \in \mathcal{P}$. Suppose there exist $C, c, s > 0$ such that*

(i) $f_P \in \mathcal{H}_s$ with $\|f_P\|_{\mathcal{H}_s} \leq C$ for all $P \in \mathcal{P}$.

(ii) Each $P \in \mathcal{P}$ is absolutely continuous with respect to Lebesgue measure on $[0,1]^d$, with corresponding density p_P satisfying

$$\sup_{P\in\mathcal{P}}\sup_{z\in[0,1]^d} p_P(z) \leq C \quad \text{and} \quad \inf_{P\in\mathcal{P}}\inf_{z\in[0,1]^d} p_P(z) \geq c.$$

(iii) $\mathrm{Var}_P(Y \mid Z) \leq C$ for all $P \in \mathcal{P}$.

If $K\log(K)/n \to 0$, then the conclusions of Proposition 3.12 hold for any $r \geq s$ with $\zeta = s/d$ and $\sigma_n^2 = C$.

Proof. Under Assumptions *(i)* and *(ii)* of the corollary, we have that Assumption *(i)* of Proposition 3.12 holds with $\zeta = s/d$ when $r \geq s$ by Propositions 3.10 and 3.11. Assumptions *(ii)* and *(iii)* of Proposition 3.12 also hold by hypothesis with $\sigma_n^2 = C$, so the conclusion follows. □

Now suppose that $\boldsymbol{\phi}^X$ and $\boldsymbol{\phi}^Z$ are the d_X- and d_Z-tensor B-spline bases of $\mathcal{S}^{d_X}_{r,N_X}$ and $\mathcal{S}^{d_Z}_{r,N_Z}$ respectively. It will be convenient to have the following decomposition of functions in the span of $\boldsymbol{\phi}^X \otimes \boldsymbol{\phi}^Z$.

Proposition 3.13. *Let $\boldsymbol{\phi} := \boldsymbol{\phi}^X \otimes \boldsymbol{\phi}^Z$, and let $K_X := (N_X + r)^{d_X}$, $K_Z := (N_Z + r)^{d_Z}$ and $K_{XZ} := K_X K_Z$. Denote by V the subspace of $R^{K_{XZ}}$ given by $V := \{\mathbf{1} \otimes \boldsymbol{v} : \boldsymbol{v} \in \mathbb{R}^{K_Z}\}$ where $\mathbf{1}$ denotes the vector of ones in $\mathbb{R}^{K_X}$. Let $V^\perp$ denote the orthogonal complement of V in $\mathbb{R}^{K_{XZ}}$, and let $\boldsymbol{\Pi} : \mathbb{R}^{K_{XZ}} \to V^\perp$ denote the projection onto $V^\perp$. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{K_{XZ}})^\top \in \mathbb{R}^{K_{XZ}}$, and define $f : [0,1]^{d_X+d_Z} \to \mathbb{R}$ by $f(x,z) := \boldsymbol{\phi}(x,z)^\top \boldsymbol{\beta}$. Then, writing $\bar{\boldsymbol{\beta}} = (\bar\beta_1, \ldots, \bar\beta_{K_Z})^\top \in \mathbb{R}^{K_Z}$, where $\bar\beta_k := K_X^{-1} \sum_{\ell=1}^{K_X} \beta_{(k-1)K_X+\ell}$, we have $(\boldsymbol{I} - \boldsymbol{\Pi})\boldsymbol{\beta} = \mathbf{1} \otimes \bar{\boldsymbol{\beta}}$ and*

$$f(x,z) = \boldsymbol{\phi}(x,z)^\top \boldsymbol{\Pi}\boldsymbol{\beta} + \boldsymbol{\phi}^Z(z)^\top \bar{\boldsymbol{\beta}}. \tag{3.66}$$

Moreover, $\|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty \le 2\|\boldsymbol{\beta}\|_\infty$.

Proof. We claim that $V^\perp = \{\boldsymbol{u} = (\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_{K_Z}^\top)^\top \in \mathbb{R}^{K_{XZ}} : \mathbf{1}^\top \boldsymbol{u}_k = 0 \ \forall k \in [K_Z]\}$. To see this, note that if $\boldsymbol{u} = (\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_{K_Z}^\top)^\top \in \mathbb{R}^{K_{XZ}}$ satisfies $\mathbf{1}^\top \boldsymbol{u}_k = 0$ for all $k \in [K_Z]$ and $\mathbf{1} \otimes \boldsymbol{v} \in V$ for some $\boldsymbol{v} = (v_1, \ldots, v_{K_Z})^\top \in \mathbb{R}^{K_Z}$, then

$$(\mathbf{1} \otimes \boldsymbol{v})^\top \boldsymbol{u} = \sum_{k=1}^{K_Z} v_k (\mathbf{1}^\top \boldsymbol{u}_k) = 0,$$

which establishes our claim. We can therefore write

$$\boldsymbol{\beta} = \mathbf{1} \otimes \bar{\boldsymbol{\beta}} + \boldsymbol{\beta} - (\mathbf{1} \otimes \bar{\boldsymbol{\beta}}),$$

where $\mathbf{1} \otimes \bar{\boldsymbol{\beta}} \in V$ and $\boldsymbol{\beta} - (\mathbf{1} \otimes \bar{\boldsymbol{\beta}}) \in V^\perp$, so $(\boldsymbol{I} - \boldsymbol{\Pi})\boldsymbol{\beta} = \mathbf{1} \otimes \bar{\boldsymbol{\beta}}$ and $\boldsymbol{\Pi}\boldsymbol{\beta} = \boldsymbol{\beta} - (\mathbf{1} \otimes \bar{\boldsymbol{\beta}})$. Hence,

$$\begin{aligned}\boldsymbol{\phi}(x,z)^\top (\boldsymbol{I} - \boldsymbol{\Pi})\boldsymbol{\beta} &= \operatorname{vec}(\boldsymbol{\phi}^X(x)\boldsymbol{\phi}^Z(z)^\top)^\top \operatorname{vec}(\mathbf{1}\bar{\boldsymbol{\beta}}^\top) = \sum_{\ell=1}^{K_X}\sum_{k=1}^{K_Z} \boldsymbol{\phi}^X_\ell(x)\boldsymbol{\phi}^Z_k(z)\bar\beta_k \\ &= \boldsymbol{\phi}^Z(z)^\top \bar{\boldsymbol{\beta}},\end{aligned}$$

by Proposition 3.9(a), from which (3.66) follows. Finally, $\|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty \le \|\boldsymbol{\beta}\|_\infty + \|\mathbf{1} \otimes \bar{\boldsymbol{\beta}}\|_\infty \le 2\|\boldsymbol{\beta}\|_\infty$. □

Our next two lemmas will be used in the proof of Proposition 3.14, which is the analogue of Corollary 3.2 for a key setting for us, namely where our response variable for spline regression consists of fitted values from an earlier spline regression.

Lemma 3.14. *Let P denote a distribution of (X,Z) on $[0,1]^{d_X} \times [0,1]^{d_Z}$ that is absolutely continuous with respect to Lebesgue measure, and let $p_{X|Z}$ denote the conditional density of X given Z. Assume that $p_{X|Z}(x\,|\,\cdot) \in \mathcal{H}^{d_Z}_s$ for every $x \in [0,1]^{d_X}$ and that there exists $C > 0$ such that*

$$\sup_{x \in [0,1]^{d_X}} \|p_{X|Z}(x\,|\,\cdot)\|_{\mathcal{H}_s} \le C.$$

Let $\boldsymbol{\phi} = (\phi_1, \dots, \phi_K)^\top$ *denote the* d_X*-tensor B-spline basis of* $\mathcal{S}^{d_X}_{r,N}$*, let* $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^\top \in \mathbb{R}^K$ *and define* $g : [0,1]^{d_Z} \to \mathbb{R}$ *by*

$$g(z) := \boldsymbol{\beta}^\top \mathbb{E}(\boldsymbol{\phi}(X) \,|\, Z = z) = \sum_{k=1}^{K} \beta_k \mathbb{E}(\phi_k(X) \,|\, Z = z).$$

Then $g \in \mathcal{H}^{d_Z}_s$ *and* $\|g\|_{\mathcal{H}_s} \leq C\|\boldsymbol{\beta}\|_\infty$.

Proof. Repeated application of Klenke (2020, Theorem 6.28) allows us to interchange derivatives and integrals such that for any multi-index $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{d_Z})^\top \in \mathbb{N}_0^{d_Z}$ with $|\boldsymbol{\alpha}| \leq \lceil s \rceil - 1 =: s_0$, we have

$$D^{\boldsymbol{\alpha}} g(z) = \sum_{k=1}^{K} \beta_k \int_{[0,1]^{d_X}} \phi_k(x) \cdot D^{\boldsymbol{\alpha}} p_{X|Z}(x \,|\, z) \,\mathrm{d}x.$$

Thus, by Hölder's inequality and the fact that the $\{\phi_k\}_{k=1}^K$ are non-negative and form a partition of unity by Proposition 3.9(a), we have

$$\|D^{\boldsymbol{\alpha}} g\|_\infty \leq \|\boldsymbol{\beta}\|_\infty \sup_{z \in [0,1]^{d_Z}} \int_{[0,1]^{d_X}} \sum_{k=1}^{K} \phi_k(x) |D^{\boldsymbol{\alpha}} p_{X|Z}(x \,|\, z)| \,\mathrm{d}x \leq C\|\boldsymbol{\beta}\|_\infty.$$

By a similar argument, when $|\boldsymbol{\alpha}| = s_0$, we have

$$\begin{aligned} |D^{\boldsymbol{\alpha}} g(z) - D^{\boldsymbol{\alpha}} g(z')| &\leq \|\boldsymbol{\beta}\|_\infty \int_{[0,1]^{d_X}} \sum_{k=1}^{K} \phi_k(x) |D^{\boldsymbol{\alpha}} p_{X|Z}(x \,|\, z) - D^{\boldsymbol{\alpha}} p_{X|Z}(x \,|\, z')| \,\mathrm{d}x \\ &\leq C\|\boldsymbol{\beta}\|_\infty \|z - z'\|_2^{s - s_0}, \end{aligned}$$

for all $z, z' \in [0,1]^{d_Z}$, as required. □

Lemma 3.15. *Let* $\{\phi_k\}_{k=1}^K$ *denote the uniform* d*-tensor B-spline basis of* $\mathcal{S}^d_{r,N}$*. For* $k \in [K]$*, suppose that* $h_k : [0,1]^d \to \mathbb{R}$ *can be written as* $h_k = s_k + r_k$ *where* $s_k \in \mathcal{S}^d_{r,N}$ *and* $r_k \in \mathcal{H}^d_s$*, and let* $M := \max_{k \in [K]} \|r_k\|_{\mathcal{H}_s}$*. Define* $m : [0,1]^d \to \mathbb{R}$ *by*

$$m(z) := \sum_{k=1}^{K} g_k(z) \phi_k(z).$$

Then there exist $C(d,r) > 0$ *and* $m^* \in \mathcal{S}^d_{2r-1,N}$ *such that*

$$\|m - m^*\|_\infty \leq \frac{MC(d,r)}{(2rK)^{\min(s,r)/d}}.$$

Proof. Let

$$\widetilde{\mathcal{S}} := \operatorname{span}(\{\phi_k(z)\phi_\ell(z)\}_{k,\ell \in [K]}) \subseteq \mathcal{S}^d_{2r-1,N}.$$

For $k \in [K]$, let r_k^* denote a supremum norm approximant to r_k in $\mathcal{S}_{r,N}^d$ (see Proposition 3.10), so that $m^* := \sum_{k=1}^K (s_k + r_k^*)\phi_k \in \widetilde{\mathcal{S}}$. Then by Hölder's inequality, Proposition 3.9(a) and Proposition 3.10, we have

$$\|m - m^*\|_\infty = \Big\| \sum_{k=1}^K (r_k - r_k^*)\phi_k \Big\|_\infty \leq \max_{k \in K} \|r_k - r_k^*\|_\infty \leq \frac{MC(d,r)}{(2rK)^{\min(s,r)/d}},$$

as desired. □

We are now in a position to state our main result on the performance of the spline on spline regression procedure.

Proposition 3.14. *Let $r \in \mathbb{N}$, let $d = d_X + d_Z$ and let $\boldsymbol{\phi}$ denote the d-tensor B-spline basis of $\mathcal{S}_{r,N}^d$. Let $\mathcal{P}$ be a family of distributions of (X, Z) on $[0,1]^{d_X} \times [0,1]^{d_Z}$. Suppose that each $P \in \mathcal{P}$ is absolutely continuous with respect to Lebesgue measure on $[0,1]^d$. Further, suppose that there exist $C, c > 0$ such that:*

(i) There exists $s \in (0, r]$ such that the conditional density $p_{X|Z,P}$ of X given Z, satisfies $p_{X|Z,P}(x|\cdot) \in \mathcal{H}_s^{d_Z}$ for every $x \in [0,1]^{d_X}$ and $\sup_{x \in [0,1]^{d_X}} \|p_{X|Z,P}(x\,|\,\cdot)\|_{\mathcal{H}_s} \leq C$.

(ii) For every $P \in \mathcal{P}$, the density $p_{Z,P}$ of Z satisfies

$$\sup_{P \in \mathcal{P}} \sup_{z \in [0,1]^d} p_{Z,P}(z) \leq C \quad \text{and} \quad \inf_{P \in \mathcal{P}} \inf_{z \in [0,1]^d} p_{Z,P}(z) \geq c.$$

Let $(X_1, Z_1), \ldots, (X_n, Z_n)$ be independent and identically distributed copies of (X, Z). For $n \in \mathbb{N}$, let $\boldsymbol{\beta} \equiv \boldsymbol{\beta}_n \in \mathbb{R}^{K_X K_Z}$ satisfy $\|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty = O(1)$ where $\boldsymbol{\Pi}$ is defined in Proposition 3.13, and define $f_n \in \mathcal{S}_{r,N}^d$ by $f_n(x,z) := \boldsymbol{\beta}^\top \boldsymbol{\phi}(x,z)$. Further, define $g_{P,n} : [0,1]^{d_Z} \to \mathbb{R}$ by $g_{P,n}(z) := \mathbb{E}_P(f_n(X,Z)\,|\,Z = z)$, let $\boldsymbol{\psi}$ denote the d_Z-tensor B-spline basis of $\mathcal{S}_{2r-1,N}^{d_Z}$ and let $\widetilde{K}_Z := (2r-1+N)^{d_Z}$. Let $Y_i := f_n(X_i, Z_i)$ for $i \in [n]$, and let $\widehat{\boldsymbol{\theta}}$ denote the ordinary least squares estimate from regressing $Y_1, \ldots, Y_n$ onto $\boldsymbol{\psi}(Z_1), \ldots, \boldsymbol{\psi}(Z_n)$. If $\widetilde{K}_Z \log(\widetilde{K}_Z)/n \to 0$, then

$$\frac{1}{n} \sum_{i=1}^n \big(g_{P,n}(Z_i) - \widehat{\boldsymbol{\theta}}^\top \boldsymbol{\psi}(Z_i)\big)^2 = O_{\mathcal{P}}(\|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty^2 \{\widetilde{K}_Z^{-2s/d_Z} + \widetilde{K}_Z/n\}).$$

Letting $\boldsymbol{\theta}_P \in \mathbb{R}^{\widetilde{K}_Z}$ be the unique solution to $g_{P,n}^\dagger(z) = \boldsymbol{\theta}_P^\top \boldsymbol{\psi}(z)$, we have under the same assumptions that

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_P\|_2^2 = O_{\mathcal{P}}(\|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty^2 \widetilde{K}_Z^2/n).$$

Finally, if (X^*, Z^*) is a new observation of (X, Z) independent of the original sample, then

$$\mathbb{E}_P\Big(\{g_{P,n}(Z^*) - \widehat{\boldsymbol{\theta}}^\top \boldsymbol{\phi}(Z^*)\}^2 \mid \widehat{\boldsymbol{\theta}}\Big) = O_{\mathcal{P}}(\|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty^2 \{\widetilde{K}_Z^{-2s/d_Z} + \widetilde{K}_Z/n\}).$$

Proof. We check the conditions of Proposition 3.12 with $\mathcal{P}$ in that result taken to be the set of distributions of (Y_1, Z_1). To this end, let $\boldsymbol{\phi}^Z$ and $\boldsymbol{\phi}^X$ denote the d_Z- and d_X-tensor B-spline

bases of $\mathcal{S}^{d_Z}_{r,N}$ and $\mathcal{S}^{d_X}_{r,N}$, respectively, so that $\boldsymbol{\phi}(x,z) = \boldsymbol{\phi}^X(x) \otimes \boldsymbol{\phi}^Z(z)$. By Proposition 3.13, we can write

$$f_n(x,z) = \boldsymbol{\phi}(x,z)^\top \boldsymbol{\Pi}\boldsymbol{\beta} + \boldsymbol{\phi}^Z(z)^\top \bar{\boldsymbol{\beta}}.$$

Thus,

$$g_{P,n}(z) = \sum_{k=1}^{K_Z} \phi_k^Z(z) \Big[\sum_{\ell=1}^{K_X} (\boldsymbol{\Pi}\boldsymbol{\beta})_{(k-1)K_X+\ell} \mathbb{E}(\phi_\ell^X(X) \,|\, Z = z) + \bar{\beta}_k \Big] =: \sum_{k=1}^{K_Z} \phi_k^Z(z) h_k(z).$$

By Lemma 3.14 and Assumption *(i)*, we have for every $k \in [K_Z]$ that the function $r_k : [0,1]^{d_Z} \to \mathbb{R}$ given by

$$r_k(z) := \sum_{\ell=1}^{K_X} (\boldsymbol{\Pi}\boldsymbol{\beta})_{(k-1)K_X+\ell} \mathbb{E}(\phi_\ell^X(X) \,|\, Z = z)$$

belongs to $\mathcal{H}_s^{d_Z}$ with $\|r_k\|_{\mathcal{H}_s} \leq C\|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty$. Since the constant function $z \mapsto \bar{\beta}_k$ belongs to $\mathcal{S}^{d_Z}_{r,N}$, we deduce from Proposition 3.11 and Lemma 3.15 that the $L_2(\mathcal{P})$-best approximant $g^\dagger_{P,n}$ to $g_{P,n}$ in $\mathcal{S}^{d_Z}_{2r-1,N}$ satisfies for each $P \in \mathcal{P}$ that

$$\begin{aligned} \|g_{P,n} - g^\dagger_{P,n}\|_\infty \leq M(C,c,d_Z,2r-1)\|g_{P,n} - g^*_{P,n}\|_\infty &\leq \frac{M(C,c,d_Z,2r-1)C\|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty}{(2rK_Z)^{s/d_Z}} \\ &\leq \frac{2^{d_Z} M(C,c,d_Z,2r-1)C\|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty}{(2r\widetilde{K_Z})^{s/d_Z}}. \end{aligned}$$

Thus, Assumption (i) of Proposition 3.12 is satisfied with $\zeta = s/d_Z$. Assumption (ii) of Proposition 3.12 is true by hypothesis, and Assumption (iii) of Proposition 3.12 holds with $\sigma_n = \|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty$ since

$$\mathrm{Var}(Y_1 \,|\, Z_1) = \mathrm{Var}(f_n(X,Z) \,|\, Z) = \mathrm{Var}(\boldsymbol{\phi}(X,Z)^\top \boldsymbol{\Pi}\boldsymbol{\beta} \,|\, Z) \leq \|\boldsymbol{\phi}(x,z)^\top \boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty^2 \leq \|\boldsymbol{\Pi}\boldsymbol{\beta}\|_\infty^2$$

by Proposition 3.9(b). The conclusions therefore follow from Proposition 3.12. □

Finally in this section, we present two results that control two different out-of-sample product errors in a sharper way than would be obtained via a naive application of the Cauchy–Schwarz inequality. The first can be regarded as a restated and uniform version of Theorem 8 and Lemma A5 in Newey and Robins (2018).

Proposition 3.15. *Let $\mathcal{P}$ be a family of distributions of (X,Y,Z) on $\mathbb{R} \times \mathbb{R} \times [0,1]^d$ with corresponding regression functions $f_P, g_P : [0,1]^d \to \mathbb{R}$ given by $f_P(z) := \mathbb{E}_P(Y \,|\, Z = z)$ and $g_P(z) := \mathbb{E}_P(X \,|\, Z = z)$ satisfying:*

(i) There exist $\zeta_f(d,r) \equiv \zeta_f > 0$ and $\zeta_g(d,r) \equiv \zeta_g > 0$ such that

$$\sup_{P\in\mathcal{P}} \|f_P - f_P^\dagger\|_\infty = O(K^{-\zeta_f}), \quad \sup_{P\in\mathcal{P}} \|g_P - g_P^\dagger\|_\infty = O(K^{-\zeta_g}),$$

where $f_P^\dagger$ and $g_P^\dagger$ denote the $L_2(P)$-best approximants of f_P and g_P respectively in $\mathcal{S}^d_{r,N}$.

(ii) Each $P \in \mathcal{P}$ is absolutely continuous with respect to Lebesgue measure on $[0,1]^d$, with corresponding density p_P satisfying

$$\sup_{P\in\mathcal{P}} \sup_{z\in[0,1]^d} p_P(z) \le C \quad \text{and} \quad \inf_{P\in\mathcal{P}} \inf_{z\in[0,1]^d} p_P(z) \ge c.$$

(iii) There exists a positive sequence $(\sigma_n^2)_{n\in\mathbb{N}}$ such that $\max\{\mathrm{Var}(Y \mid Z), \mathrm{Var}(X \mid Z)\} \le \sigma_n^2 = O(1)$.

Now suppose we are given three independent samples $(X_i^f, Y_i^f, Z_i^f)_{i=1}^n$, $(X_i^g, Y_i^g, Z_i^g)_{i=1}^n$ and $(X_i, Y_i, Z_i)_{i=1}^n$, each consisting of n independent and identically distributed copies of (X, Y, Z). Let $\boldsymbol{\phi}$ denote the d-tensor B-spline basis of $\mathcal{S}_{r,N}^d$. Let $\widehat{\boldsymbol{\beta}}_f$ and $\widehat{\boldsymbol{\beta}}_g$ denote the ordinary least squares estimates from regressing $Y_1^f, \dots, Y_n^f$ onto $\boldsymbol{\phi}(Z_1^f), \dots, \boldsymbol{\phi}(Z_n^f)$ and $X_1^g, \dots, X_n^g$ onto $\boldsymbol{\phi}(Z_1^g), \dots, \boldsymbol{\phi}(Z_n^g)$ respectively. Define fitted regression functions $\widehat{f}$ and $\widehat{g}$ by $\widehat{f}(z) = \widehat{\boldsymbol{\beta}}_f^\top \boldsymbol{\phi}(z)$ and $\widehat{g}(z) = \widehat{\boldsymbol{\beta}}_g^\top \boldsymbol{\phi}(z)$ respectively. If $K\log(K)/n \to 0$, then

$$\frac{1}{n}\sum_{i=1}^n \{\widehat{f}(Z_i) - f_P(Z_i)\}\{\widehat{g}(Z_i) - g_P(Z_i)\} = O_{\mathcal{P}}\Big(K^{-(\zeta_f+\zeta_g)} + \frac{K^{1/2}}{n} + \frac{K^{2-\max(\zeta_f,\zeta_g)}\log K}{n^2}\Big).$$

Proof. Suppose without loss of generality that $\zeta_f \ge \zeta_g$. Define $\boldsymbol{\beta}_{P,f}, \boldsymbol{\beta}_{P,g} \in \mathbb{R}^K$ so that $f_P^\dagger(z) = \boldsymbol{\beta}_{P,f}^\top \boldsymbol{\phi}(z)$ and $g_P^\dagger(z) = \boldsymbol{\beta}_{P,g}^\top \boldsymbol{\phi}(z)$. We start with the decomposition

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \{\widehat{f}(Z_i) - f_P(Z_i)\}\{\widehat{g}(Z_i) - g_P(Z_i)\} &= \underbrace{\frac{1}{n}\sum_{i=1}^n \{f_P^\dagger(Z_i) - f_P(Z_i)\}\{g_P^\dagger(Z_i) - g_P(Z_i)\}}_{\mathrm{I}_n} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^n \{\widehat{f}(Z_i) - f_P^\dagger(Z_i)\}\{g_P^\dagger(Z_i) - g_P(Z_i)\}}_{\mathrm{II}_n^f} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^n \{f_P^\dagger(Z_i) - f_P(Z_i)\}\{\widehat{g}(Z_i) - g_P^\dagger(Z_i)\}}_{\mathrm{II}_n^g} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^n \{\widehat{f}(Z_i) - f_P^\dagger(Z_i)\}\{\widehat{g}(Z_i) - g_P^\dagger(Z_i)\}}_{\mathrm{III}_n}.
\end{aligned}$$

By Assumption *(i)*,

$$\sup_{P\in\mathcal{P}} |\mathrm{I}_n| \le \sup_{P\in\mathcal{P}} \|f_P - f_P^\dagger\|_\infty \|g_P - g_P^\dagger\|_\infty = O\big(K^{-(\zeta_f+\zeta_g)}\big).$$

Using (3.62) in the proof of Proposition 3.12 (with g_P and $g_P^\dagger$ in place of f_P and $f_P^\dagger$ there), we deduce that

$$\begin{aligned}\mathbb{E}_P((\mathrm{II}_n^f)^2 \mid \widehat{f}) &= \frac{1}{n}(\widehat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}_{P,f})^\top \mathbb{E}_P(\{g_P^\dagger(Z) - g_P(Z)\}^2 \boldsymbol{\phi}(Z)\boldsymbol{\phi}(Z)^\top)(\widehat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}_{P,f}) \\ &\leq \frac{1}{n}\|g_P^\dagger - g_P\|_\infty^2 \|\boldsymbol{\Sigma}_P\|_{\mathrm{op}} \|\widehat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}_{P,f}\|_2^2 = O_{\mathcal{P}}(K^{-(2\zeta_g - 1)} n^{-2}),\end{aligned}$$

by our hypothesis on $\|g_P^\dagger - g_P\|_\infty$, Proposition 3.9(d) and Proposition 3.12. Thus, by Lemma 3.2,

$$\mathrm{II}_n^f = O_{\mathcal{P}}(K^{-(\zeta_g - 1/2)} n^{-1}) = O_{\mathcal{P}}(K^{1/2} n^{-1}).$$

Similarly,

$$\mathrm{II}_n^g = O_{\mathcal{P}}(K^{1/2} n^{-1}).$$

To deal with the III_n term, define $\widehat{\boldsymbol{\Sigma}} := n^{-1}\sum_{i=1}^n \boldsymbol{\phi}(Z_i)\boldsymbol{\phi}(Z_i)^\top$, $\widehat{\boldsymbol{\Sigma}}_f := n^{-1}\sum_{i=1}^n \boldsymbol{\phi}(Z_i^f)\boldsymbol{\phi}(Z_i^f)^\top$ and $\widehat{\boldsymbol{\Sigma}}_g := n^{-1}\sum_{i=1}^n \boldsymbol{\phi}(Z_i^g)\boldsymbol{\phi}(Z_i^g)^\top$. For $i \in [n]$, let $\varepsilon_i^f := Y_i^f - f_P(Z_i^f)$, $\varepsilon_i^g := X_i^g - g_P(Z_i^g)$, $h_i^f := f_P(Z_i^f) - f_P^\dagger(Z_i^f)$ and $h_i^g := g_P(Z_i^g) - g_P^\dagger(Z_i^g)$. We write

$$\begin{aligned}\mathrm{III}_n = (\widehat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}_{P,f})^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{P,g}) &= \underbrace{\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^f \boldsymbol{\phi}(Z_i^f)^\top\Big)\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{P,g})}_{\mathrm{III}_n^{(1)}} \\ &+ \underbrace{\Big(\frac{1}{n}\sum_{i=1}^n h_i^f \boldsymbol{\phi}(Z_i^f)^\top\Big)\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}_g^{-1}\Big(\frac{1}{n}\sum_{i=1}^n h_i^g \boldsymbol{\phi}(Z_i^g)\Big)}_{\mathrm{III}_n^{(2)}} \\ &+ \underbrace{\Big(\frac{1}{n}\sum_{i=1}^n h_i^f \boldsymbol{\phi}(Z_i^f)^\top\Big)\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}_g^{-1}\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^g \boldsymbol{\phi}(Z_i^g)\Big)}_{\mathrm{III}_n^{(3)}}.\end{aligned}$$

To deal with the $\mathrm{III}_n^{(1)}$ term, we have using the fact that $\mathbb{E}_P(\varepsilon_1^f \mid Z_1^f) = 0$ and $\mathrm{Var}_P(Y_1^f \mid Z_1^f) = \mathbb{E}((\varepsilon_1^f)^2 \mid Z_1^f) \leq \sigma_n^2$ that

$$\begin{aligned}\mathbb{E}_P((\mathrm{III}_n^{(1)})^2 \mid \widehat{\boldsymbol{\beta}}_g, (Z_i, Z_i^f)_{i=1}^n) &\leq \frac{\sigma_n^2}{n}(\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{P,g})^\top \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{P,g}) \\ &\leq \frac{\sigma_n^2}{n}\|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}^2 \|\widehat{\boldsymbol{\Sigma}}_f^{-1}\|_{\mathrm{op}} \|\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{P,g}\|_2^2 = O_{\mathcal{P}}(Kn^{-2}),\end{aligned}$$

by our hypothesis on σ_n^2, (3.60) in Lemma 3.13 and Proposition 3.12. Hence, by another application of Lemma 3.2,

$$\mathrm{III}_n^{(1)} = O_{\mathcal{P}}(K^{1/2} n^{-1}).$$

To deal with the $\mathrm{III}_n^{(2)}$ term, by the Cauchy–Schwarz inequality,

$$|\mathrm{III}_n^{(2)}| \leq \|\widehat{\boldsymbol{\Sigma}}_f^{-1/2}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}_g^{-1/2}\|_{\mathrm{op}} \Big\|\widehat{\boldsymbol{\Sigma}}_f^{-1/2}\frac{1}{n}\sum_{i=1}^n h_i^f \boldsymbol{\phi}(Z_i^f)\Big\|_2 \Big\|\widehat{\boldsymbol{\Sigma}}_g^{-1/2}\frac{1}{n}\sum_{i=1}^n h_i^g \boldsymbol{\phi}(Z_i^g)\Big\|_2.$$

By (3.60) in Lemma 3.13,

$$\|\widehat{\mathbf{\Sigma}}_f^{-1/2}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}_g^{-1/2}\|_{\mathrm{op}} = O_{\mathcal{P}}(1).$$

The same argument as in (3.64) in the proof of Proposition 3.12 now yields that

$$\mathrm{III}_n^{(2)} = O_{\mathcal{P}}(K^{-(\zeta_f+\zeta_g-1)}n^{-1}) = O_{\mathcal{P}}(K^{-(\zeta_f+\zeta_g)}).$$

To deal with the $\mathrm{III}_n^{(3)}$ term we write

$$\begin{aligned}\mathrm{III}_n^{(3)} = {} & \underbrace{\Big(\frac{1}{n}\sum_{i=1}^n h_i^f \boldsymbol{\phi}(Z_i^f)^\top\Big)\mathbf{\Sigma}_P^{-1}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}_g^{-1}\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^g\boldsymbol{\phi}(Z_i^g)\Big)}_{\mathrm{III}_n^{(3,1)}} \\ & + \underbrace{\Big(\frac{1}{n}\sum_{i=1}^n h_i^f \boldsymbol{\phi}(Z_i^f)^\top\Big)(\widehat{\mathbf{\Sigma}}_f^{-1}-\mathbf{\Sigma}_P^{-1})\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}_P^{-1}\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^g\boldsymbol{\phi}(Z_i^g)\Big)}_{\mathrm{III}_n^{(3,2)}} \\ & + \underbrace{\Big(\frac{1}{n}\sum_{i=1}^n h_i^f \boldsymbol{\phi}(Z_i^f)^\top\Big)(\widehat{\mathbf{\Sigma}}_f^{-1}-\mathbf{\Sigma}_P^{-1})\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{\Sigma}}_g^{-1}-\mathbf{\Sigma}_P^{-1})\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^g\boldsymbol{\phi}(Z_i^g)\Big)}_{\mathrm{III}_n^{(3,3)}}.\end{aligned}$$

For the first term, we have by an argument similar to the $\mathrm{III}_n^{(1)}$ term that

$$\mathbb{E}_P\big((\mathrm{III}_n^{(3,1)})^2 \,|\, (Z_i, Z_i^f, Z_i^g)_{i=1}^n\big) \le \frac{\sigma_n^2}{n}\Big\|\frac{1}{n}\sum_{i=1}^n h_i^f\boldsymbol{\phi}(Z_i^f)\Big\|_2^2 \|\mathbf{\Sigma}_P^{-1}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}_g^{-1}\widehat{\mathbf{\Sigma}}_g\widehat{\mathbf{\Sigma}}_g^{-1}\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}}.$$

By Lemma 3.2, our assumption on σ_n, (3.63) in the proof of Proposition 3.12, Proposition 3.9(d) and (3.60) in the proof of Lemma 3.13, we therefore have

$$\mathrm{III}_n^{(3,1)} = O_{\mathcal{P}}(K^{1/2-\zeta_f}n^{-1}) = O_{\mathcal{P}}(K^{1/2}n^{-1}).$$

Similarly,

$$\mathbb{E}_P\big((\mathrm{III}_n^{(3,2)})^2 \,|\, (Z_i, Z_i^f, Z_i^g)_{i=1}^n\big) \le \frac{\sigma_n^2}{n}\Big\|\frac{1}{n}\sum_{i=1}^n h_i^f\boldsymbol{\phi}(Z_i^f)\Big\|_2^2 \|\mathbf{\Sigma}_P^{-1}-\widehat{\mathbf{\Sigma}}_f^{-1}\|_{\mathrm{op}}^2\|\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}_P^{-1}\widehat{\mathbf{\Sigma}}_g\mathbf{\Sigma}_P^{-1}\widehat{\mathbf{\Sigma}}\|_{\mathrm{op}}.$$

Hence, by the same arguments as for $\mathrm{III}_n^{(3,1)}$, together with the second result in Lemma 3.13,

$$\mathrm{III}_n^{(3,2)} = O_{\mathcal{P}}\Big(\frac{K^{-(\zeta_f-1)}\log^{1/2}(eK)}{n^{3/2}}\Big) = O_{\mathcal{P}}(K^{1/2}n^{-1}).$$

Finally, by the Cauchy–Schwarz inequality, we have

$$|\mathrm{III}_n^{(3,3)}| \le \|\widehat{\mathbf{\Sigma}}\|_{\mathrm{op}}\|\widehat{\mathbf{\Sigma}}_f^{-1}-\mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}}\|\widehat{\mathbf{\Sigma}}_g^{-1}-\mathbf{\Sigma}_P^{-1}\|_{\mathrm{op}}\Big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i^g\boldsymbol{\phi}(Z_i^g)\Big\|_2\Big\|\frac{1}{n}\sum_{i=1}^n h_i^f\boldsymbol{\phi}(Z_i^f)\Big\|_2,$$

so

$$\mathrm{III}_n^{(3,3)} = O_{\mathcal{P}}\left(\frac{K^{2-\zeta_f}\log(eK)}{n^2}\right) = O_{\mathcal{P}}\left(K^{-(\zeta_f+\zeta_g)} + \frac{K^{2-\max(\zeta_f,\zeta_g)}\log K}{n^2}\right)$$

from our previous bounds. The result follows. □

Our second and final result controls a different type of product error and is loosely based on Theorem 8 and Corollary 9 in a working version of Ichimura and Newey (2015).

Proposition 3.16. *Let $\mathcal{P}$ be a family of distributions of (Y, Z) on $\mathbb{R}\times[0,1]^d$ with corresponding regression function $f_P : [0,1]^d \to \mathbb{R}$ given by $f_P(z) := \mathbb{E}_P(Y \mid Z = z)$. Further, let $(g_P)_{P\in\mathcal{P}}$ be a family of functions from $[0,1]^d$ to $\mathbb{R}$ with $\rho_P := \mathbb{E}_P(g_P(Z)^2) < \infty$ and $\inf_{P\in\mathcal{P}} \rho_P > 0$. Suppose that*

(i) There exist $\zeta_f(d,r) \equiv \zeta_f > 0$ and $\zeta_g(d,r) \equiv \zeta_g > 0$ such that

$$\sup_{P\in\mathcal{P}} \|f_P - f_P^\dagger\|_\infty = O(K^{-\zeta_f}), \quad \sup_{P\in\mathcal{P}} \|g_P - g_P^\dagger\|_\infty = O(K^{-\zeta_g}),$$

where $f_P^\dagger$ and $g_P^\dagger$ denote the $L_2(P)$-best approximants of f_P and g_P respectively in $\mathcal{S}_{r,N}^d$.

(ii) Each $P \in \mathcal{P}$ is absolutely continuous with respect to Lebesgue measure on $[0,1]^d$, with corresponding density p_P satisfying

$$\sup_{P\in\mathcal{P}} \sup_{z\in[0,1]^d} p_P(z) \le C \quad \text{and} \quad \inf_{P\in\mathcal{P}} \inf_{z\in[0,1]^d} p_P(z) \ge c.$$

(iii) There exists a positive sequence $(\sigma_n^2)_{n\in\mathbb{N}}$ such that $\mathrm{Var}(Y \mid Z) \le \sigma_n^2 = O(1)$.

Now suppose we are given two independent samples $(Y_i^f, Z_i^f)_{i=1}^n$ and $(Y_i, Z_i)_{i=1}^n$, each consisting of n independent and identically distributed copies of (Y, Z). Let $\boldsymbol{\phi}$ denote the d-tensor B-spline basis of $\mathcal{S}_{r,N}^d$. Let $\widehat{\boldsymbol{\beta}}$ denote the ordinary least squares estimate from regressing $Y_1^f, \dots, Y_n^f$ onto $\boldsymbol{\phi}(Z_1^f), \dots, \boldsymbol{\phi}(Z_n^f)$. Define the fitted regression function $\widehat{f}$ by $\widehat{f}(z) = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{\phi}(z)$. If $K\log(K)/n \to 0$, then

$$\frac{1}{n}\sum_{i=1}^n g_P(Z_i)\{\widehat{f}(Z_i) - f_P(Z_i)\} = O_{\mathcal{P}}(K^{-(\zeta_f+\zeta_g)} + K^{-(\zeta_g-1/2)}n^{-1} + \rho_P^{1/2}n^{-1/2}\{1 + K^{-(\zeta_f-1/2)}\}).$$

Proof. Define $\boldsymbol{\beta}_{P,f}, \boldsymbol{\beta}_{P,g} \in \mathbb{R}^K$ so that $f_P^\dagger(z) = \boldsymbol{\beta}_{P,f}^\top\boldsymbol{\phi}(z)$ and $g_P^\dagger(z) = \boldsymbol{\beta}_{P,g}^\top\boldsymbol{\phi}(z)$. By Proposition 3.9(b),

$$\|\boldsymbol{\beta}_{P,g}\|_2 \le K^{1/2}c_s(r)^{-d}\{\|g_P^\dagger - g_P\|_\infty + \|g_P\|_2\} \le K^{1/2}c_s(r)^{-d}\{\|g_P^\dagger - g_P\|_\infty + \rho_P^{1/2}c^{-1/2}\}.$$

Thus, by *(i)*,

$$\|\boldsymbol{\beta}_{P,g}\|_2 = O_{\mathcal{P}}(K^{-(\zeta_g-1/2)} + \rho_P^{1/2}K^{1/2}). \tag{3.67}$$

We can write

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} g_P(Z_i)\{\widehat{f}(Z_i) - f_P(Z_i)\} \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\{g_P(Z_i) - g_P^\dagger(Z_i)\}\{f_P^\dagger(Z_i) - f_P(Z_i)\}}_{\mathrm{I}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} g_P^\dagger(Z_i)\{f_P^\dagger(Z_i) - f_P(Z_i)\}}_{\mathrm{II}_n} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\{g_P(Z_i) - g_P^\dagger(Z_i)\}\{\widehat{f}(Z_i) - f_P^\dagger(Z_i)\}}_{\mathrm{III}_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} g_P^\dagger(Z_i)\{\widehat{f}(Z_i) - f_P^\dagger(Z_i)\}}_{\mathrm{IV}_n}.
\end{aligned}$$

By assumption *(i)* again,

$$\sup_{P\in\mathcal{P}} |\mathrm{I}_n| \le \sup_{P\in\mathcal{P}} \|f_P - f_P^\dagger\|_\infty \|g_P - g_P^\dagger\|_\infty = O\big(K^{-(\zeta_f+\zeta_g)}\big).$$

From (3.62) in the proof of Proposition 3.12,

$$\mathbb{E}_P\big(g_P^\dagger(Z)\{f_P(Z) - f_P^\dagger(Z)\}\big) = \mathbb{E}_P\big(\{f_P(Z) - f_P^\dagger(Z)\}\boldsymbol{\phi}(Z)^\top \boldsymbol{\beta}_{P,g}\big) = 0,$$

and therefore

$$\mathbb{E}_P(\mathrm{II}_n^2) = \frac{1}{n}\boldsymbol{\beta}_{P,g}^\top \mathbb{E}_P\big(\{f_P(Z) - f_P^\dagger(Z)\}^2 \boldsymbol{\phi}(Z)\boldsymbol{\phi}(Z)^\top\big)\boldsymbol{\beta}_{P,g} \le \frac{1}{n}\|f_P - f_P^\dagger\|_\infty^2 \|\boldsymbol{\beta}_{P,g}\|_2^2 \|\boldsymbol{\Sigma}_P\|_{\mathrm{op}}.$$

Thus,

$$\mathrm{II}_n = O_{\mathcal{P}}\big(K^{-(\zeta_f-\zeta_g)} n^{-1/2} + \rho_P^{1/2} K^{-\zeta_f} n^{-1/2}\big) = O_{\mathcal{P}}\big(K^{-\zeta_f-\zeta_g} + \rho_P^{1/2} n^{-1/2}\big).$$

The same argument as for the II_n^f term in the proof of Proposition 3.15 yields that

$$\mathrm{III}_n = O_{\mathcal{P}}\big(K^{-(\zeta_g - 1/2)} n^{-1}\big).$$

To deal with IV_n, we write, with quantities defined as in the proof of Proposition 3.15,

$$\mathrm{IV}_n = \boldsymbol{\beta}_{P,g}^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{P,f}) = \underbrace{\boldsymbol{\beta}_{P,g}^\top \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}_f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^f \boldsymbol{\phi}(Z_i^f)\right)}_{\mathrm{IV}_n^{(1)}} + \underbrace{\boldsymbol{\beta}_{P,g}^\top \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}_f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} h_i^f \boldsymbol{\phi}(Z_i^f)\right)}_{\mathrm{IV}_n^{(2)}}.$$

Since $\mathbb{E}_P(\varepsilon_1^f \mid Z_1^f) = 0$ and $\mathrm{Var}_P(Y_1^f \mid Z_1^f) = \mathbb{E}_P\big((\varepsilon_1^f)^2 \mid Z_1^f\big) \le \sigma_n^2$, we have

$$\mathbb{E}_P\big((\mathrm{IV}_n^{(1)})^2 \mid (Z_i, Z_i^f)_{i=1}^n\big) \le \frac{\sigma_n^2}{n}\boldsymbol{\beta}_{P,g}^\top \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_{P,g} \le \frac{\sigma_n^2}{n}\|\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\Sigma}}_f\widehat{\boldsymbol{\Sigma}}_f^{-1}\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}\|\boldsymbol{\beta}_{P,g}\|_2^2,$$

so

$$\mathrm{IV}_n^{(1)} = O_{\mathcal{P}}\big((K^{-\zeta_g} + \rho_P^{1/2})n^{-1/2}\big) = O_{\mathcal{P}}\big(\rho_P^{1/2} n^{-1/2}\big).$$

Finally, by the Cauchy–Schwarz inequality,

$$\begin{aligned}|\mathrm{IV}_n^{(2)}| &\leq \|\boldsymbol{\beta}_{P,g}\|_2 \|\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}_f^{-1}\|_{\mathrm{op}} \Big\| \frac{1}{n}\sum_{i=1}^n h_i^f \boldsymbol{\phi}(Z_i^f) \Big\|_2 \\ &= O_{\mathcal{P}}(K^{-(\zeta_f+\zeta_g)+1/2} n^{-1/2} + \rho_P^{1/2} K^{-(\zeta_f-1/2)} n^{-1/2}) \\ &= O_{\mathcal{P}}(K^{-(\zeta_f+\zeta_g)} + \rho_P^{1/2} K^{-(\zeta_f-1/2)} n^{-1/2}).\end{aligned}$$

The result follows. □

3.11 Univariate linear model analysis

In this section we give a more detailed analysis of the setting considered in Section 3.3.1. In contrast to the remainder of this chapter, we let $\mathcal{D}_1$ contain n_1 observations and $\mathcal{D}_2$ contain n_2 observations, and we let $2n = n_1 + n_2$ for this subsection only. All limiting statements in this section are interpreted as $\min\{n_1, n_2\} \to 0$. This will facilitate a discussion of the effect of the splitting ratio on the power, and to compare the power of the proposed test more precisely with existing methods. To simplify our analysis, we set $\widehat{v} \equiv 1$. We now formally write down the assumption required for the main result of this section (Proposition 3.17).

Assumption 3.6. Suppose that the family $\mathcal{P}$ of joint distributions P of (X, Y, Z) satisfies the linear model (3.8). Let $\boldsymbol{\eta}_P$ and $\boldsymbol{\theta}_P$ denote the population least squares projections of X on Z and Y on Z, respectively. Let $\widehat{\beta}$, $\widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\theta}}$ denote estimators of β_P, $\boldsymbol{\eta}_P$ and $\boldsymbol{\theta}_P$, respectively, where $\widehat{\beta}$ is trained on $\mathcal{D}_2$ and $\widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\theta}}$ are trained on $\mathcal{D}_1$. Writing $\sigma^2_{\beta_P} := \mathrm{Var}_P(\sqrt{n_2}(\widehat{\beta} - \beta_P))$, suppose that $\widehat{\beta}$, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\eta}}$ satisfy

$$\sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}} \Big| \mathbb{P}_P(\sqrt{n_2}\sigma_{\beta_P}^{-1}(\widehat{\beta} - \beta_P) \leq t) - \Phi(t) \Big| = o(1), \tag{3.68}$$

$$\sqrt{n_1}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_P\|_2 \cdot \Big\| \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} (X_i - \boldsymbol{\eta}_P^\top Z_i) Z_i \Big\|_2 = o_{\mathcal{P}}(1), \tag{3.69}$$

$$\sqrt{n_1}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_P\|_2 \cdot \Big\| \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} (Y_i - \boldsymbol{\theta}_P^\top Z_i) Z_i \Big\|_2 = o_{\mathcal{P}}(1), \tag{3.70}$$

$$\sqrt{n_1}\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_P\|_2 \cdot \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_P\|_2 \cdot \Big\| \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} Z_i Z_i^\top \Big\|_{\mathrm{op}} = o_{\mathcal{P}}(1), \tag{3.71}$$

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_P\|_2^2 \cdot \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} (X_i - \boldsymbol{\eta}_P^\top Z_i)^2 \|Z_i\|_2^2 = o_{\mathcal{P}}(1), \tag{3.72}$$

$$\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_P\|_2^2 \cdot \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} (Y_i - \boldsymbol{\theta}_P^\top Z_i)^2 \|Z_i\|_2^2 = o_{\mathcal{P}}(1), \tag{3.73}$$

$$\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_P\|_2^2 \cdot \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_P\|_2^2 \cdot \frac{1}{n_1}\sum_{i\in\mathcal{I}_1} \|Z_i\|_2^4 = o_{\mathcal{P}}(1). \tag{3.74}$$

In Section 3.3.1 we show consider a simpler but less general assumption (Assumption 3.1) that suffices for the analysis when the estimators are OLS estimators. These more general

assumptions allow for settings where alternate estimators are used, or the dimension is allowed to increase with n.

As mentioned before, we set the estimated weight function $\widehat{v} \equiv 1$ in this analysis, which yields $L_i = \widehat{\beta}(Y_i - \widehat{\boldsymbol{\theta}}^\top Z_i)(X_i - \widehat{\boldsymbol{\eta}}^\top Z_i) =: \widehat{\beta} R_i$ for $i = 1, \dots, n_1$. The resulting PCM statistic is

$$T = \operatorname{sgn}(\widehat{\beta}) \frac{\frac{1}{\sqrt{n_1}} \sum_{i \in \mathcal{I}_1} R_i}{\sqrt{\frac{1}{n_1} \sum_{i \in \mathcal{I}_1} R_i^2 - \left(\frac{1}{n_1} \sum_{i \in \mathcal{I}_1} R_i\right)^2}}.$$

To simplify our presentation, we write $\sigma^2_{P,X|Z} := \mathbb{E}_P\left(\{X - \boldsymbol{\eta}_P^\top Z\}^2\right)$ and $\sigma^2_{P,XY|Z} := \operatorname{Var}_P(\{Y - \boldsymbol{\theta}_P^\top Z\}\{X - \boldsymbol{\eta}_P^\top Z\})$. The following result provides asymptotic size and power expressions for the PCM test in this context.

Proposition 3.17. *Suppose that $\mathcal{P}$ is a family of distributions P of (X, Y, Z) for which the estimators $\widehat{\beta}$, $\widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\theta}}$ satisfy Assumption 3.6. In addition, assume that there exist $c, C, \delta > 0$ such that $\sigma^2_{P,XY|Z} > c$, $\mathbb{E}_P\{|(Y - \boldsymbol{\theta}_P^\top Z)(X - \boldsymbol{\eta}_P^\top Z)|^{2+\delta}\} \leq C$ and $\sqrt{n_1}|\beta_P|\sigma^2_{P,X|Z} \leq C$ for all $P \in \mathcal{P}$ and $n \in \mathbb{N}$. Then, by letting*

$$\psi_{P,\alpha,n} := \Phi\left(\frac{\sqrt{n_2}\beta_P}{\sigma_{\beta_P}}\right) \cdot \Phi\left(z_\alpha + \frac{\sqrt{n_1}\beta_P \sigma^2_{P,X|Z}}{\sigma_{P,XY|Z}}\right) + \Phi\left(-\frac{\sqrt{n_2}\beta_P}{\sigma_{\beta_P}}\right) \cdot \Phi\left(z_\alpha - \frac{\sqrt{n_1}\beta_P \sigma^2_{P,X|Z}}{\sigma_{P,XY|Z}}\right),$$

the power of the PCM test satisfies

$$\sup_{P \in \mathcal{P}} \left|\mathbb{P}_P(T > z_{1-\alpha}) - \psi_{P,\alpha,n}\right| \to 0, \text{ as } \min\{n_1, n_2\} \to \infty.$$

Furthermore, when $\alpha < 1/2$, we have $\psi_{P,\alpha,n} \geq \alpha$ and, when $\sigma^2_{P,X|Z}/\sigma_{P,XY|Z} > 0$, equality holds if and only if $\beta_P = 0$.

Proposition 3.17 confirms that under Assumption 3.6 and the given moment conditions, our proposed test is asymptotically valid uniformly over the null hypothesis $\mathcal{P}_0 := \{P \in \mathcal{P} : \beta_P = 0\}$. In terms of splitting ratio, a consequence of Proposition 3.17, as stated formally in Corollary 3.1 is that in this linear model setting one cannot hope to achieve high power against a local alternative where $\tau_P \asymp n^{-1}$ unless $n_1 \asymp n_2$. While limited to the linear model, we think this result instils confidence in our choice of balanced splitting ratio, and also reveals that the choice of splitting ratio that maximises the asymptotic power depends on the underlying (unknown) parameters. For this reason, we consider $n_1 = n_2$ by default for simplicity.

For the specific class of linear alternatives considered in Proposition 3.17, the asymptotic power of the GCM test (Shah and Peters, 2020) without sample splitting is

$$\Phi\left(z_{\alpha/2} + \frac{\sqrt{n_1 + n_2}\beta_P \sigma^2_{P,X|Z}}{\sigma_{P,XY|Z}}\right) + \Phi\left(z_{\alpha/2} - \frac{\sqrt{n_1 + n_2}\beta_P \sigma^2_{P,X|Z}}{\sigma_{P,XY|Z}}\right).$$

Comparing this expression with $\psi_{P,\alpha,n}$, one can see that the GCM test is typically more powerful than the proposed test, but only by a constant factor when $n_1 \asymp n_2$. However, as mentioned earlier, the proposed test can have power against broader alternatives than the GCM

test depending on the choice of projection. In comparison with the tests of Williamson et al. (2022) and Dai et al. (2021), the proposed test achieves higher power. In particular, their tests become powerless whenever $\sqrt{n}\tau_P \to 0$, which is true for both parametric and nonparametric settings. Moreover, as pointed out by Williamson et al. (2022) and further demonstrated in Appendix 3.9.2, their tests via sample splitting may not control the Type I error when (X, Y, Z) are mutually independent. In contrast, our approach does not suffer from this issue and can be powerful even when $\sqrt{n}\tau_P \to 0$, as demonstrated by Proposition 3.17. In the next subsection, we provide the proof of Proposition 3.17.

3.11.1 Proof of Proposition 3.17

As $L_i = \widehat{\beta}(Y_i - \widehat{\boldsymbol{\theta}}^\top Z_i)(X_i - \widehat{\boldsymbol{\eta}}^\top Z_i) =: \widehat{\beta} R_i$ for $i = 1, \ldots, n_1$, recall that our test statistic is

$$T_n = \operatorname{sgn}(\widehat{\beta}) \frac{\frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} R_i}{\sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1} R_i^2 - \left(\frac{1}{n_1}\sum_{i=1}^{n_1} R_i\right)^2}}. \tag{3.75}$$

Let

$$T_{R,n} := \frac{\frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} (R_i - \beta \sigma^2_{X|Z})}{\sigma_{XY|Z}},$$

and for now suppose that the following approximations hold:

$$\sup_{P \in \mathcal{P}} \sup_{t \in \mathbb{R}} \left| \mathbb{P}(T_{R,n} \leq t) - \Phi(t) \right| = o(1) \tag{3.76}$$

and

$$\left\{ \frac{\sigma^2_{XY|Z}}{\frac{1}{n_1}\sum_{i=1}^{n_1} R_i^2 - \left(\frac{1}{n_1}\sum_{i=1}^{n_1} R_i\right)^2} \right\}^{1/2} = 1 + o_{\mathcal{P}}(1). \tag{3.77}$$

Then, by (3.77), together with the hypotheses that $\sqrt{n_1}|\beta|\sigma^2_{X|Z} \leq C$ and $\sigma^2_{XY|Z} > c$, we have by Lemma 3.3, a uniform version of Slutsky's theorem (Bengs and Holzmann, 2019, Theorem 6.3) that

$$\begin{aligned} T_n &= \left\{ \operatorname{sgn}(\widehat{\beta}) T_{R,n} + \operatorname{sgn}(\widehat{\beta}) s_{n_1,\beta} \right\} \{1 + o_{\mathcal{P}}(1)\} \\ &= \operatorname{sgn}(\widehat{\beta}) T_{R,n} + \operatorname{sgn}(\widehat{\beta}) s_{n_1,\beta} + V_n, \end{aligned}$$

where $V_n = o_{\mathcal{P}}(1)$. Define $s_{n_1,\beta} := \sqrt{n_1}\beta\sigma^2_{X|Z}/\sigma_{XY|Z}$ and note that, since $(R_i)_{i=1}^{n_1}$ and $\operatorname{sgn}(\widehat{\beta})$ are formed on independent data and are thus independent, we have

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} |\mathbb{P}(T_n > z_{1-\alpha}) - \psi_{\alpha,n}| = \sup_{P\in\mathcal{P}} \left|\mathbb{P}\left(\operatorname{sgn}(\widehat{\beta})T_{R,n} > z_{1-\alpha} - \operatorname{sgn}(\widehat{\beta})s_{n_1,\beta} - V_n\right) - \psi_{\alpha,n}\right| \\
&\leq \sup_{P\in\mathcal{P}} \left|\mathbb{P}(\operatorname{sgn}(\widehat{\beta}) = 1)\mathbb{P}\left(T_{R,n} > z_{1-\alpha} - s_{n_1,\beta} - V_n\right) - \Phi\left(\frac{\sqrt{n_2}\beta}{\sigma_\beta}\right)\Phi\left(z_\alpha + s_{n_1,\beta}\right)\right| \\
&\qquad + \sup_{P\in\mathcal{P}} \left|\mathbb{P}(\operatorname{sgn}(\widehat{\beta}) = -1)\mathbb{P}\left(-T_{R,n} > z_{1-\alpha} + s_{n_1,\beta} - V_n\right) - \Phi\left(\frac{-\sqrt{n_2}\beta}{\sigma_{\beta_P}}\right)\Phi\left(z_\alpha - s_{n_1,\beta}\right)\right|.
\end{aligned}$$

Both of the two terms in this upper bound are dealt with similarly, so we only show how to argue that the first term is $o(1)$. To this end, we have

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \left|\mathbb{P}(\operatorname{sgn}(\widehat{\beta}) = 1)\mathbb{P}\left(T_{R,n} > z_{1-\alpha} - s_{n_1,\beta} - V_n\right) - \Phi\left(\frac{\sqrt{n_2}\beta}{\sigma_\beta}\right)\Phi\left(z_\alpha + s_{n_1,\beta}\right)\right| \\
&\leq \sup_{P\in\mathcal{P}} \left|\mathbb{P}(\operatorname{sgn}(\widehat{\beta}) = 1) - \Phi\left(\frac{\sqrt{n_2}\beta}{\sigma_\beta}\right)\right| + \sup_{P\in\mathcal{P}} \left|\mathbb{P}\left(T_{R,n} > z_{1-\alpha} - s_{n_1,\beta} - V_n\right) - \Phi\left(z_\alpha + s_{n_1,\beta}\right)\right|.
\end{aligned}$$

The first term is $o(1)$ by (3.68). To deal with the second term, we write

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \left|\mathbb{P}\left(T_{R,n} > z_{1-\alpha} - s_{n_1,\beta} - V_n\right) - \Phi\left(z_\alpha + s_{n_1,\beta}\right)\right| \\
&\qquad \leq \sup_{P\in\mathcal{P}} \left|\mathbb{P}\left(T_{R,n} > z_{1-\alpha} - s_{n_1,\beta} - V_n\right) - 1 + \Phi\left(z_{1-\alpha} - s_{n_1,\beta} - V_n\right)\right| \\
&\qquad\qquad + \sup_{P\in\mathcal{P}} \left|1 - \Phi\left(z_{1-\alpha} - s_{n_1,\beta} - V_n\right) - \Phi\left(z_\alpha + s_{n_1,\beta}\right)\right|,
\end{aligned}$$

and note that (3.76) implies that the first term is $o(1)$. For the second term, observe by the symmetry of the standard Gaussian that if $W \sim N(0,1)$ and $\epsilon > 0$, then

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \left|\Phi\left(z_{1-\alpha} - s_{n_1,\beta} - V_n\right) - \Phi\left(z_{1-\alpha} - s_{n_1,\beta}\right)\right| \\
&\qquad \leq \sup_{P\in\mathcal{P}} \mathbb{P}\left(\left|W - z_{1-\alpha} + s_{n_1,\beta}\right| \leq |V_n|\right) \\
&\qquad \leq \sup_{P\in\mathcal{P}} \mathbb{P}\left(\left|W - z_{1-\alpha} + \frac{\sqrt{n_1}\beta\sigma^2_{X|Z}}{\sigma_{P,XY|Z}}\right| \leq \epsilon\right) + \sup_{P\in\mathcal{P}} \mathbb{P}(|V_n| \geq \epsilon) \\
&\qquad \leq \frac{2\epsilon}{\sqrt{2\pi}} + o(1).
\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, the first claim of the proposition will follow once we establish (3.76) and (3.77).

For the claim (3.76), consider the decomposition

$$T_{R,n} = \underbrace{\frac{\sigma_{XY|Z}^{-1}}{\sqrt{n_1}} \sum_{i=1}^{n_1} \{(Y_i - \boldsymbol{\theta}^\top Z_i)(X_i - \boldsymbol{\eta}^\top Z_i) - \beta\sigma_{X|Z}^2\}}_{T_n^{(1)}}$$
$$- \underbrace{\frac{\sigma_{XY|Z}^{-1}}{\sqrt{n_1}} \sum_{i=1}^{n_1} (Y_i - \boldsymbol{\theta}^\top Z_i)(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^\top Z_i}_{T_n^{(2)}} - \underbrace{\frac{\sigma_{XY|Z}^{-1}}{\sqrt{n_1}} \sum_{i=1}^{n_1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top Z_i (X_i - \boldsymbol{\eta}^\top Z_i)}_{T_n^{(3)}}$$
$$+ \underbrace{\frac{\sigma_{XY|Z}^{-1}}{\sqrt{n_1}} \sum_{i=1}^{n_1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top Z_i Z_i^\top (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})}_{T_n^{(4)}}.$$

By the assumption that $\mathbb{E}\{|(Y - \boldsymbol{\theta}^\top Z)(X - \boldsymbol{\eta}^\top Z)|^{2+\delta}\} \leq C$, Shah and Peters (2020, Lemma 18) yields that

$$\sup_{P \in \mathcal{P}} \sup_{t \in \mathbb{R}} |\mathbb{P}(T_n^{(1)} \leq t) - \Phi(t)| \to 0.$$

Moreover,

$$|T_n^{(2)}| \leq c\sqrt{n_1} \|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 \Big\| \frac{1}{n_1} \sum_{i=1}^{n_1} (Y_i - \boldsymbol{\theta}^\top Z_i) Z_i \Big\|_2 = o_{\mathcal{P}}(1),$$

by Cauchy–Schwarz, the assumption that $\sigma_{P,XY|Z} > c$ and (3.70). We can argue similarly that $T_n^{(3)} = o_{\mathcal{P}}(1)$ using (3.69). Finally,

$$|T_n^{(4)}| \leq c\sqrt{n_1} \|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 \Big\| \frac{1}{n_1} \sum_{i=1}^{n_1} Z_i Z_i^\top \Big\|_{\mathrm{op}} = o_{\mathcal{P}}(1)$$

by similar arguments as above and (3.71). Combining the above with the uniform version of Slutsky's theorem, we have the desired claim (3.76).

To prove (3.77), we let $\widetilde{R}_{n,i} := R_{n,i} - \beta\sigma_{X|Z}^2$ for $i \in [n_1]$ and note that

$$\frac{1}{n_1} \sum_{i=1}^{n_1} R_i^2 - \Big(\frac{1}{n_1} \sum_{i=1}^{n_1} R_i \Big)^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} \widetilde{R}_{n,i}^2 - \Big(\frac{1}{n_1} \sum_{i=1}^{n_1} \widetilde{R}_{n,i} \Big)^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} \widetilde{R}_{n,i}^2 + o_{\mathcal{P}}(1),$$

where the second equality follows from the proof of (3.76) above. To ease the notation further, for $i \in [n_1]$, we write

$$\widetilde{R}_{n,i} = \underbrace{(Y_i - \boldsymbol{\theta}^\top Z_i)(X_i - \boldsymbol{\eta}^\top Z_i) - \beta\sigma_{X|Z}^2}_{\mathrm{I}_i} - \underbrace{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top Z_i (X_i - \boldsymbol{\eta}^\top Z_i)}_{\mathrm{II}_i}$$
$$- \underbrace{(Y_i - \boldsymbol{\theta}^\top Z_i)(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})^\top Z_i}_{\mathrm{III}_i} + \underbrace{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top Z_i Z_i^\top (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})}_{\mathrm{IV}_i}.$$

Then

$$\begin{aligned}\frac{1}{n_1}\sum_{i=1}^{n_1}\widetilde{R}_{n,i}^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathrm{I}_i^2 + \frac{1}{n_1}\sum_{i=1}^{n_1}\mathrm{II}_i^2 + \frac{1}{n_1}\sum_{i=1}^{n_1}\mathrm{III}_i^2 + \frac{1}{n_1}\sum_{i=1}^{n_1}\mathrm{IV}_i^2 \\ - \frac{2}{n_1}\sum_{i=1}^{n_1}\mathrm{I}_i\mathrm{II}_i - \frac{2}{n_1}\sum_{i=1}^{n_1}\mathrm{I}_i\mathrm{III}_i + \frac{2}{n_1}\sum_{i=1}^{n_1}\mathrm{I}_i\mathrm{IV}_i \\ + \frac{2}{n_1}\sum_{i=1}^{n_1}\mathrm{II}_i\mathrm{III}_i - \frac{2}{n_1}\sum_{i=1}^{n_1}\mathrm{II}_i\mathrm{IV}_i - \frac{2}{n_1}\sum_{i=1}^{n_1}\mathrm{III}_i\mathrm{IV}_i.\end{aligned}$$

By the assumption that $\mathbb{E}\{|(Y-\boldsymbol{\theta}^\top Z)(X-\boldsymbol{\eta}^\top Z)|^{2+\delta}\} \le C$, Shah and Peters (2020, Lemma 19) yields that $\sigma_{XY|Z}^{-2} n_1^{-1}\sum_{i=1}^{n_1}\mathrm{I}_i^2 = 1 + o_{\mathcal{P}}(1)$. Moreover, by Cauchy–Schwarz,

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\mathrm{II}_i^2 \le \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_2^2 \cdot \frac{1}{n_1}\sum_{i=1}^{n_1}(X_i-\boldsymbol{\eta}^\top Z_i)^2\|Z_i\|_2^2,$$

so (3.72) together with $\sigma_{XY|Z}^2 > c$ implies that $\frac{\sigma_{XY|Z}^{-2}}{n_1}\sum_{i=1}^{n_1}\mathrm{II}_i^2 = o_{\mathcal{P}}(1)$. Similarly,

$$\frac{\sigma_{XY|Z}^{-2}}{n_1}\sum_{i=1}^{n_1}\mathrm{III}_i^2 = o_{\mathcal{P}}(1)$$

by (3.73). By two applications of Cauchy–Schwarz, we have

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\mathrm{IV}_i^2 \le \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_2^2\|\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}\|_2^2 \cdot \frac{1}{n_1}\sum_{i=1}^{n_1}\|Z_i\|_2^4,$$

so (3.74) combined with the lower bound on $\sigma_{XY|Z}^2$ yields that $\frac{\sigma_{XY|Z}^{-2}}{n_1}\sum_{i=1}^{n_1}\mathrm{IV}_i^2 = o_{\mathcal{P}}(1)$. Turning to the cross-product terms, by Cauchy–Schwarz and the previous analysis,

$$\left|\frac{\sigma_{XY|Z}^{-2}}{n_1}\sum_{i=1}^{n_1}\mathrm{I}_i\mathrm{II}_i\right| \le \sigma_{XY|Z}^{-2}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\mathrm{I}_i^2\right)^{1/2}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\mathrm{II}_i^2\right)^{1/2} = o_{\mathcal{P}}(1).$$

The other terms can be similarly analysed and shown to be $o_{\mathcal{P}}(1)$. We have thus established by the uniform version of Slutsky's theorem that

$$\frac{1}{\sigma_{XY|Z}^2}\left\{\frac{1}{n_1}\sum_{i=1}^{n_1}R_i^2 - \left(\frac{1}{n_1}\sum_{i=1}^{n_1}R_i\right)^2\right\} = 1 + o_{\mathcal{P}}(1).$$

Finally, (3.77) follows by the above result combined with Lemma 3.7. This completes the proof of the first claim in Proposition 3.17.

To prove the second claim, let us assume that $\beta \geq 0$ (the case $\beta < 0$ can be handled very similarly), and denote

$$\begin{aligned}\psi_{\alpha,n} &= \underbrace{\Phi\Big(\frac{\sqrt{n_2}\beta}{\sigma_\beta}\Big)}_{V(\beta)} \cdot \underbrace{\Phi\Big(z_\alpha + s_{n_1,\beta}\Big)}_{W_1(\beta)} + \underbrace{\Phi\Big(-\frac{\sqrt{n_2}\beta}{\sigma_\beta}\Big)}_{1-V(\beta)} \cdot \underbrace{\Phi\Big(z_\alpha - s_{n_1,\beta}\Big)}_{W_2(\beta)} \\ &= W_2(\beta) + V(\beta)\{W_1(\beta) - W_2(\beta)\}.\end{aligned}$$

Then $V(\beta) \geq 1/2$ and $W_1(\beta) - W_2(\beta) \geq 0$, so $\psi_{\alpha,n} \geq W_2(\beta) + \{W_1(\beta) - W_2(\beta)\}/2 = W_1(\beta)/2 + W_2(\beta)/2$.

Next observe that the function $\delta \mapsto \Phi(z_\alpha + \delta)/2 + \Phi(z_\alpha - \delta)/2$ is continuous on $\mathbb{R}$, and when $\alpha < 1/2$, it is decreasing when $\delta < 0$ and increasing when $\delta > 0$. It follows that

$$\psi_{\alpha,n} \geq \frac{1}{2}\Phi\Big(z_\alpha + s_{n_1,\beta}\Big) + \frac{1}{2}\Phi\Big(z_\alpha - s_{n_1,\beta}\Big) \geq \alpha,$$

and when $\sigma^2_{X|Z}/\sigma_{XY|Z} > 0$, we have equality in both inequalities if and only if $\beta = 0$.

3.12 Additional simulation results

3.12.1 Linear model comparison

To compare the local power properties of the PCM with the approach considered in Williamson et al. (2021) and the more conventional F-test with robust standard error (White, 1980) (as implemented in the R package `lmtest` (Zeileis and Hothorn, 2002)), we consider the following setup where Z and ξ are independent $N_5(0, \boldsymbol{I})$ random vectors, $\varepsilon \sim N(0,1)$ independently of Z and ξ and

$$\begin{aligned}X &= Z + \xi, \\ Y &= \boldsymbol{\beta}^\top X + 2\Big(\big(1 + e^{-3X_1}\big)^{-1} + \big(1 + e^{-3Z_1}\big)^{-1}\Big)\varepsilon.\end{aligned}$$

We simulate $n \in \{100, 400, 1600, 6400\}$ observations from the above model. All regressions are performed using OLS, except $\widehat{v}$ which uses a random forest and we only apply Algorithm 2 for simplicity (rather than doing multiple sample splits).

The results can be seen in Figure 3.4. They confirm our theoretical observations in Section 3.9.2 on the power properties of the PCM and `williamson` tests in this linear model setting. It is also interesting to note that, except for the smallest sample size, here the PCM has greater power than `lmtest` due to our modelling of the heteroscedasticity in the data.

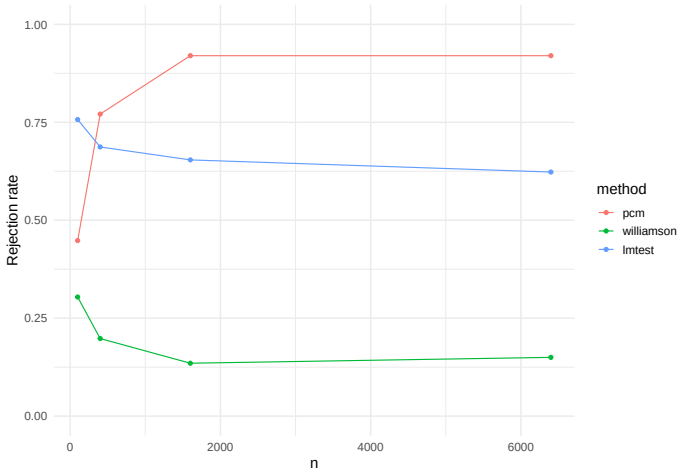


Fig. 3.4 Power in the alternative settings considered in Section 3.12.1 for nominal 5%-level tests.

3.12.2 Generalised additive models with binary responses

Here we consider settings similar to those considered in Section 3.6.1, but with Y binary. Our null settings use

$$\mathbb{P}(Y = 1) = \text{logit}(\sin(2\pi Z_1)),$$

and we consider three alternative settings mirroring those in Section 3.6.1:

1. $\mathbb{P}(Y = 1) = \text{logit}(\sin(2\pi Z_1) + 0.25X^2)$,
2. $\mathbb{P}(Y = 1) = \text{logit}(\sin(2\pi Z_1) + 0.5X^2)$,
3. $\mathbb{P}(Y = 1) = \text{logit}(\sin(2\pi Z_1) + 0.5X^2 Z_2)$.

For all regressions with binary responses, we fit a binomial generalised additive model with logistic link, and we use additive models for all other regressions; we use the implementations in the R-package `mgcv` (Wood, 2017). The results can be seen in Figure 3.5 and are broadly in line with those in Section 3.6.1 with the PCM performing favourably though being powerless in Setting 3 with pure interactions (as to be expected), and `williamson` and most notably `gam` not maintaining Type I error control.

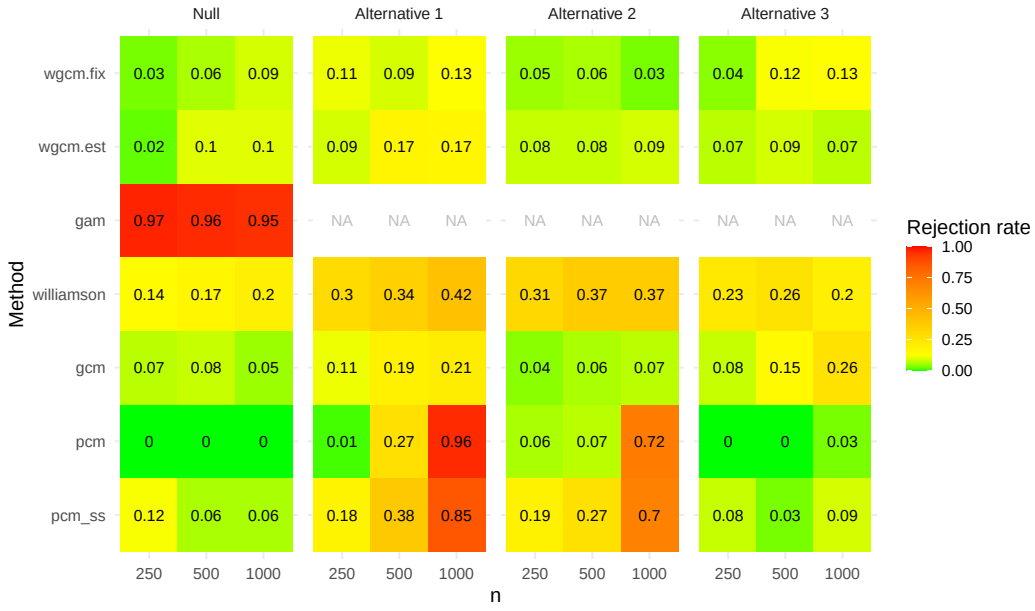


Fig. 3.5 Rejection rates in the various settings considered in Section 3.12.2 for nominal 5%-level tests.

References

Aït-Sahalia, Y., Bickel, P. J., and Stoker, T. M. (2001). Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *Journal of Econometrics*, 105(2):363–412.

Arias-Castro, E., Pelletier, B., and Saligrama, V. (2018). Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30(2):448–471.

Arnold, S. (1984). The asymptotic validity of invariant procedures for the repeated measures model and multivariate linear model. *Journal of Multivariate Analysis*, 15(3):325–335.

Bahadur, R. R. and Savage, L. J. (1956). The Nonexistence of Certain Statistical Procedures in Nonparametric Problems. *The Annals of Mathematical Statistics*, 27(4):1115 – 1122.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Balakrishnan, S. and Wasserman, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727–749.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577 – 606.

Barber, R. F., Candès, E. J., and Samworth, R. J. (2020). Robust inference with knockoffs. *The Annals of Statistics*, 48(3):1409 – 1431.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

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

Bengs, V. and Holzmann, H. (2019). Uniform approximation in classical weak convergence theory.

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

Berrett, T. B., Kontoyiannis, I., and Samworth, R. J. (2021). Optimal rates for independence testing via U-statistic permutation tests. *The Annals of Statistics*, 49(5):2457 – 2490.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197.

Billingsley, P. (1999). *Convergence of Probability Measures*. John Wiley & Sons, Inc.

Bogachev, V. (2018). *Weak Convergence of Measures*. Mathematical Surveys and Monographs. American Mathematical Society.

Bogachev, V. I. (2007). *Measure Theory*. Springer Berlin Heidelberg.

Bojer, C. S. and Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37(2):587–603.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

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

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition.

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

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Canonne, C. L., Diakonikolas, I., Kane, D. M., and Stewart, A. (2018). Testing conditional independence of discrete distributions. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–57.

Carothers, N. L. (2000). *Real Analysis*. Cambridge University Press.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SigKDD international conference on knowledge discovery and data mining*, pages 785–794.

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

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

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

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4):417–451.

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

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

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

Dai, B., Shen, X., and Pan, W. (2021). Significance tests of feature relevance for a blackbox learner. *arXiv preprint arXiv:2103.04985 (to appear in IEEE Transactions on Neural Networks and Learning Systems)*.

de Boor, C. (1976). Splines as linear combinations of b-splines. a survey.

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

Diakonikolas, I. and Kane, D. M. (2016). A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE.

DiCiccio, C. J., DiCiccio, T. J., and Romano, J. P. (2020). Exact tests via multiple data splitting. *Statistics & Probability Letters*, 166:108865.

Doob, J. L. (1994). *Measure Theory*. Springer New York.

Duchesne, P. and 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.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition.

Durrett, R. (2019). *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition.

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

Fan, Y. and Li, Q. (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica: Journal of the econometric society*, 64(4):865–890.

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 (Applied Statistics)*, 33(3):332–339.

Fernández, T. and Rivera, N. (2022). A general framework for the analysis of kernel-based tests. *arXiv preprint arXiv:2209.00124*.

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

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer New York.

Fisher, R. A. (1924). The distribution of the partial correlation coefficient. *Metron*, 3:329–332.

Fugarolas, M. and Cobos, F. (1983). On schauder bases in the lorentz operator ideal. *Journal of Mathematical Analysis and Applications*, 95(1):235–242.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20(4):830–851. PMID: 22368438.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. T. (2020). *refund: Regression with Functional Data*. R package version 0.1-22.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv för matematik*, 1(3):195–277.

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

Gut, A. (2013). *Probability: A Graduate Course*. Springer New York.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer New York.

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

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

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Heinze-Deml, C., Peters, J., and Meinshausen, N. (2018). Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016.

Helwig, N. E. (2018). *eegkit: Toolkit for Electroencephalography Data*. R package version 1.0-4.

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

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, Ltd.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600 – 1635.

Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799 – 821.

Ichimura, H. and Newey, W. K. (2015). The influence function of semiparametric estimators. *arXiv preprint arXiv:1508.01378*.

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 indicatorsof electroencephalography. *Phys. Rev. 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.

Ingster, Y. I. (1982). On the minimax nonparametric detection of signals in white gaussian noise. *Problemy Peredachi Informatsii*, 18:61–73.

Ingster, Y. I. (1987). Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337.

Ingster, Y. I. (2000). Adaptive chi-square tests. *Journal of Mathematical Sciences*, 99(2):1110–1119.

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

Janková, J., Shah, R. D., Bühlmann, P., and Samworth, R. J. (2020). Goodness-of-fit testing in high dimensional generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):773–795.

Jin, Z., Yan, X., and Matteson, D. S. (2018). Testing for Conditional Mean Independence with Covariates through Martingale Difference Divergence. *arXiv preprint arXiv:1805.06640*.

Kasy, M. (2019). Uniformity and the delta method. *Journal of Econometric Methods*, 8(1):1–19.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

Kim, I., Neykov, M., Balakrishnan, S., and Wasserman, L. (2021). Local permutation tests for conditional independence. *arXiv preprint arXiv:2112.11666*.

Kim, I. and Ramdas, A. (2020). Dimension-agnostic inference using cross U-statistics. *arXiv preprint arXiv:2011.05068*.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Klenke, A. (2020). *Probability Theory*. Springer International Publishing.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Kraft, C. (1955). *Some Conditions for Consistency and Uniform Consistency of Statistical Procedures*. University of California Press.

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

Lavergne, P. and Vuong, Q. (2000). Nonparametric significance testing. *Econometric Theory*, 16(4):576–601.

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1(1):38–53.

Lepskiĭ, O. V. (1991). Asymptotically minimax adaptive estimation i: Upper bounds. optimally adaptive estimates. *Theory Probab. Appl.*, 36:682–697.

Li, D. and Queffélec, H. (2017). *Introduction to Banach Spaces: Analysis and Probability*, volume 1 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press.

Liu, H., Tang, Y., and 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., and Peters, J. (2021a). Conditional Independence Testing in Hilbert Spaces with Applications to Functional Data Analysis. *arXiv preprint arXiv:2101.07108*.

Lundborg, A. R., Shah, R. D., and Peters, J. (2021b). *ghcm: Functional Conditional Independence Testing with the GHCM*. R package version 1.0.0.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, 58(303):690–700.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, 22(4):719–748.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

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

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Neykov, M., Balakrishnan, S., and Wasserman, L. (2021). Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51.

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

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. 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., and 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., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

Petersen, L. and Hansen, N. R. (2021). Testing conditional independence via quantile regression based partial copulas. *Journal of Machine Learning Research*, 22(70):1–47.

Powell, M. J. D. (1981). *Approximation Theory and Method*. Cambridge University Press.

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

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

Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, 47(4):379–396.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer New York.

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17.

Reiss, P. T., Goldsmith, J., Shang, H. L., and Ogden, R. T. (2016). Methods for scalar-on-function regression. *International Statistical Review*, 85(2):228–249.

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

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

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(1):233–243.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469.

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

Romano, J. P. (2004). On Non-parametric Testing, the Uniform Behaviour of the t-test, and Related Problems. *Scandinavian Journal of Statistics*, 31(4):567–584.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rønn-Nielsen, A. and Hansen, E. (2014). *Conditioning and Markov properties*. Department of Mathematical Sciences, University of Copenhagen.

Scalora, F. S. (1961). Abstract martingale convergence theorems. *Pacific Journal of Mathematics*, 11(1):347–374.

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

Scheidegger, C., Hörrmann, J., and Bühlmann, P. (2021). The Weighted Generalised Covariance Measure. *arXiv preprint arXiv:2111.04361*.

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

Schilling, R. L. (2017). *Measures, Integrals and Martingales*. Cambridge University Press.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge Mathematical Library. Cambridge University Press, 3 edition.

Seal, H. L. (1967). Studies in the history of probability and statistics. xv: The historical development of the gauss linear model. *Biometrika*, 54(1/2):1–24.

Shah, R. D. and Bühlmann, P. (2018). Goodness-of-fit tests for high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):113–135.

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

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318.

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., Meek, C., Cooper, G., and Richardson, T. (2000). *Causation, Prediction, and Search*. Adaptive computation and machine learning. MIT Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Ullah, S. and Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13(1):43.

Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vakhania, N. N., Tarieladze, V. I., and Chobanyan, S. A. (1987). *Probability Distributions on Banach Spaces*. Springer Netherlands.

Valiant, G. and Valiant, P. (2017). An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Verdinelli, I. and Wasserman, L. (2021). Decorrelated Variable Importance. *arXiv preprint arXiv:2111.10853*.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4):791–808.

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

Wang, Y. and Shah, R. D. (2020). Debiased Inverse Propensity Score Weighting for Estimation of Average Treatment Effects with High-Dimensional Confounders. *arXiv preprint arXiv:2011.08661*.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178–2201.

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

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77:9–22.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2022+). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association, to appear*.

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*. Chapman and Hall/CRC.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

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

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

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

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

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

Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, 2(3):7–10.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-Based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813, Arlington, Virginia, USA. AUAI Press.

Zhang, L. and Janson, L. (2020). Floodgate: inference for model-free variable importance. *arXiv preprint arXiv:2007.01283*.

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

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