Statistical inference for the slope parameter in functional linear regression

Tim Kutta, Gauthier Dierickx, Holger Dette

August 20, 2021

Abstract

In this paper we consider the linear regression model $Y = SX + \varepsilon$ with functional regressors and responses. We develop new inference tools to quantify deviations of the true slope S from a hypothesized operator S_0 with respect to the Hilbert–Schmidt norm $|||S-S_0|||^2$, as well as the prediction error $\mathbb{E}\|SX - S_0X\|^2$. Our analysis is applicable to functional time series and based on asymptotically pivotal statistics. This makes it particularly user friendly, because it avoids the choice of tuning parameters inherent in long-run variance estimation or bootstrap of dependent data. We also discuss two sample problems as well as change point detection. Finite sample properties are investigated by means of a simulation study.

Mathematically our approach is based on a sequential version of the popular spectral cut-off estimator $\hat{S}_N$ for S. It is well-known that the L^2-minimax rates in the functional regression model, both in estimation and prediction, are substantially slower than $1/\sqrt{N}$ (where N denotes the sample size) and that standard estimators for S do not converge weakly to non-degenerate limits. However, we demonstrate that simple plug-in estimators - such as $|||\hat{S}_N - S_0|||^2$ for $|||S - S_0|||^2$ - are $\sqrt{N}$-consistent and its sequential versions satisfy weak invariance principles. These results are based on the smoothing effect of L^2-norms and established by a new proof-technique, the *smoothness shift*, which has potential applications in other statistical inverse problems.

Keywords: functional linear regression, inverse problem, relevant tests, spectral cut-off, prediction error
AMS subject classification: 62R10, 62M20

1 Introduction

Over the past decades new branches of statistical research have developed to meet the needs of an economy with growing data volumes at its disposal. One approach to analyze large data samples, particularly when detected on a dense grid, is interpolation of discrete measurements to continuous, functional observations. This method is known as functional data analysis (FDA) and nowadays has numerous applications as diverse as economics, climatology and medicine (see, for example, Andersson and Lillestøl, 2010; Bonner et al., 2014; Sørensen et al., 2013, among many others). FDA benefits users in several ways: From a

theoretical perspective functional models - in contrast to standard multivariate analysis - can incorporate smoothness in the data. From a computational viewpoint, the interpolation of thousands of observations to, say a yearly curve of price development, can drastically reduce the amount of data to be stored, since interpolations only consume a fraction of memory compared to the noisy raw data (see, for example, Liebl, 2013; Stöhr et al., 2021, among many others). From a practical point of view, random curves are easy to visualize and interpret for human users, who cannot possibly make sense from endless data lists.
One model that has attracted particular interest in the context of FDA due to its parsimony and interpretability is the functional linear regression model

$$Y_n = SX_n + \varepsilon_n \qquad n = 1, ..., N. \tag{1.1}$$

Here the regressors, errors and dependent variables are functions. More mathematically, X_n, ε_n and Y_n are elements of (potentially different) Hilbert spaces H_1 and H_2 and the slope parameter $S : H_1 \to H_2$ is a Hilbert–Schmidt operator. Such models extend existing ones for time series and panel data, and have applications in different situations, where standard, non-functional approaches fail (see, for example, Andersson and Lillestøl, 2010).
Linear models are attractive for users because of their simple structure, where all information is stored in the slope parameter S. However, compared to the better known case of finite dimensional, linear regression, the standard tasks of estimation, prediction and statistical inference become substantially more difficult in the functional regime. Indeed all of these tasks require the approximate inversion of the compact covariance operator $\Gamma := \mathbb{E}X \otimes X$ (we discuss this issue in detail in Section 2.2), which constitutes a *statistical inverse problem.* Statistical inverse problems extend *classical inverse problem* (the reconstruction of an entity using indirect observations, see fore example Engl et al., 1996) by including noise in the model. Naturally arising in different settings, statistical inverse problems have been studied extensively in tomography, deconvolution or the heat equation, to name but a few examples (see Cavalier, 2008; Bissantz and Holzmann, 2008, and the references therein). Since our subsequent discussion is exclusively concerned with statistical inverse problems, we will for brevity just refer to them as inverse problems without qualifications. Characteristic of inverse problems is a need for regularization, which leads to slower than parametric convergence rates of the resulting estimates.
The study of functional linear regression and the associated inverse problem has been a part of FDA for more than two decades (see, for example the monograph of Ramsay and Silverman, 1997). Early work has focused on the scalar response model

$$Y_n = \int_0^1 \varphi_S(t) X_n(t) dt + \varepsilon_n \ , \tag{1.2}$$

which is a special case of (1.1), where $H_1 = L^2[0,1]$, the space of square integrable function defined on the interval $[0,1]$, $H_2 = \mathbb{R}$ and $S : L^2[0,1] \to \mathbb{R}$ is an integral operator with square integrable kernel φ_S. For the investigation of scalar response models, we refer the interested reader for instance to Cardot et al. (2003), Hall and Horowitz (2007), Yuan and Cai (2012) and references therein. In Hall and Horowitz (2007) convergence rates for spectral cut-off estimators (a specific kind of regularization also used in this paper) are investigated with respect to the L^2-norm and it is shown that these estimators can achieve

minimax optimal rates. Generalizations of these results to functional linear regression with functional responses can be found in Imaizumi and Kato (2018). Similarly, Benatia et al. (2017) investigate minimax L^2-rates, as well as practical aspects of Tikhonov regularizations in the estimation of the slope parameter. Besides L^2-rates other aspects of model (1.1) have been investigated in a wide variety of works, such as consistency under weak dependence in a white noise model (see Hörmann and Kidzinski, 2012), aspects of identifiability (see Scheipl and Greven, 2016), minimax rates for prediction (see Crambes and Mas, 2013) and robust estimation (see Shin and Lee, 2016).

The list of cited references is by no means complete and only comprises a fraction of the larger body of research in this domain. Besides estimation and prediction, hypothesis testing in the functional regression model has attracted some attention. Cardot et al. (2003, 2004) consider the problem of testing for a particular value of the slope, i.e. $H_0 : S = S_0$ vs. $H_1 : S \neq S_0$, where S_0 is some hypothesized operator (see Section 3.1 for details and more references). It turns out that H_0 can be examined by $\sqrt{N}$-consistent tests, which employ transformed versions of both operators S and S_0. Importantly such tests do not have to solve the inverse problem of reconstructing S, which makes them theoretically more parsimonious, but practically somewhat difficult to interpret, as they do not assess the deviation of the true operators of interest (for a more detailed discussion of this problem we refer to Section 3.1 below). As a consequence attention has recently shifted to inference methods, based on direct slope comparisons, to make the results statistically more meaningful. For example, in the functional linear model (1.2) with scalar responses Imaizumi and Kato (2019) develop confidence bands that cover the slope function at most points with a prespecified probability. Other authors use Gaussian approximations to construct uniform confidence bands; see, for example, Babii (2020), who devises honest confidence bands for the regression function in a nonparametric instrumental variable regression using Tikhonov regularization. Notice that these approaches, based on reconstructing S (or φ_S), have to pay the price of solving the inverse problem, by a convergence speed significantly slower than $1/\sqrt{N}$.

In this paper, we contribute to the discussion by providing a new method of statistical inference in the regression model (1.1). Our inference concerns the two deviation measures $|||S - S_0|||^2$ (the distance in Hilbert–Schmidt norm) and $\mathbb{E}\|SX - S_0X\|^2$ (the expected prediction error), where again S_0 is a hypothesized operator. In contrast to the hypothesis of the form $H_0 : S = S_0$, we prefer a more quantitative approach, testing whether the deviation $|||S-S_0|||^2$ or $\mathbb{E}\|SX - S_0X\|^2$ is smaller than some predetermined threshold, say $\Delta > 0$. Although one has to solve the inverse problem to reconstruct S, the proposed estimates converge at a parametric rate of $1/\sqrt{N}$, due to a natural smoothing effect of the L^2-norms. In particular, we use a new proof-technique, called *smoothness shift*, to establish asymptotic normality for estimators of the deviation measures $|||S - S_0|||^2$ and $\mathbb{E}\|SX - S_0X\|^2$. This technique can be also used in the study of other inverse problems, such as deconvolution or tomography and is therefore of independent interest.

A direct application of these results for statistical inference such as the construction of hypothesis tests or confidence intervals is theoretically possible, but practically difficult, because it requires the estimation of asymptotic (long-run) variances. This estimation is intricate in inverse problems even for i.i.d. data and becomes even more difficult for functional time series (see Proposition 3.3 below for a presentation of the long-run variance τ^2). To circumvent these problems we investigate sequential versions of our

estimators, prove weak invariance principles and use the concept of self-normalization (see, for example, Shao, 2015; Dette et al., 2020) to construct (asymptotically) pivotal estimates of the deviation measures. Users benefit from the principle of self-normalization, because it provides (robust) inference tools, which do not require the choice of tuning parameters for long-run variance estimation (see, e.g. Horváth et al. (2011) and Kokoszka (2012)) or for the block bootstrap of dependent data (see, e.g., Politis and Romano, 1994; Bücher and Kojadinovic, 2013).
The rest of this paper is organized as follows: In Section 2, we discuss the linear model in detail and construct the spectral cut-off estimator $\hat{S}_N$ for S. Next, in Section 3, we present statistical inference for the distance in Hilbert–Schmidt norm and in Section 4 inference for the expected prediction error. Then, in Section 5 we propose extensions of our methodology to two sample and change point scenarios, while in Section 6 we investigate finite sample properties by virtue of a simulation study. Finally, the Appendix contains the technical proofs and mathematical details.

2 Estimation of the slope parameter

In this Section we introduce the mathematical set-up for estimation in the functional linear regression model (1.1). We begin by recalling some basic facts about Hilbert–Schmidt operators and continue with a discussion of the estimation problem of the slope S in the functional linear model. In particular, we explain the necessity of regularization and discuss the ensuing variance-regularization trade-off.

2.1 Operators on Hilbert spaces

Throughout this paper we treat functional observations as elements of Hilbert spaces. Thus before we proceed to the statistics, we recall some fundamental aspects of operator theory on Hilbert spaces. For a more detailed overview we recommend the monographs of Horváth and Kokoszka (2012) (with particular emphasis on functional data) as well as Weidmann (1980).
Suppose two generic Hilbert spaces $(\mathcal{H}_1, \langle , \rangle_1)$ and $(\mathcal{H}_2, \langle , \rangle_2)$ are given. The corresponding norms on the spaces are denoted by $\|\cdot\|_i$, for $i = 1, 2$. The space $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ consists of all linear operators $L : \mathcal{H}_1 \to \mathcal{H}_2$, satisfying

$$|||L|||_{\mathcal{L}} := \sup_{\|x\|_1=1} \|Lx\|_2 < \infty.$$

The norm $|||\cdot|||_{\mathcal{L}}$ is referred to as operator or spectral norm. Recall that all operators with bounded spectral norm are also continuous. An important subclass of $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ is given by the compact operators, i.e., such operators L which map the unit ball in $\mathcal{H}_1$ to a totally bounded set in $\mathcal{H}_2$. In the special case where $\mathcal{H}_1 = \mathcal{H}_2$ and the operator L is both compact and symmetric, L can - according to the spectral theorem for normal operators - be diagonalized, in the sense that for any $x \in \mathcal{H}_1$

$$Lx = \sum_{n\in\mathbb{N}} \eta_n \langle f_n, x\rangle_1 f_n, \tag{2.1}$$

where $\eta_1, \eta_2, ... \in \mathbb{R}$ are the eigenvalues and $f_1, f_2, ... \in \mathcal{H}_1$ the corresponding eigenvectors of L. In the context of functional spaces the eigenvectors are usually referred to as eigenfunctions. The most restrictive

class of operators, that we consider in this paper consists of the Hilbert–Schmidt operators. This subspace of $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ denoted by $\mathcal{S}(\mathcal{H}_1, \mathcal{H}_2)$ consists of all linear operators L, which satisfy

$$|||L||| := \sum_{n\in\mathbb{N}} \|Lf_n\|_2^2 < \infty,$$

where $\{f_n\}_{n\in\mathbb{N}}$ is some orthonormal basis (ONB) of $\mathcal{H}_1$. The value of the norm is independent of the basis. Notice that $|||L||| < \infty$ directly entails compactness and hence boundedness w.r.t. the operator norm. The norm $|||\cdot|||$ is called Hilbert–Schmidt norm and is the infinite-dimensional analogue of the Frobenius norm. Just as the Frobenius norm it is induced by an inner product, which for two operators $L, T \in \mathcal{S}(\mathcal{H}_1, \mathcal{H}_2)$ is given by

$$\langle L, T\rangle := \sum_{n\in\mathbb{N}} \langle Lf_n, Tf_n\rangle_2,$$

where again the value of the inner product (on the left hand side) does not depend on the choice of basis. Equipped with this inner product the linear space $\mathcal{S}(\mathcal{H}_1, \mathcal{H}_2)$ becomes itself a Hilbert space. Finally we introduce the outer product of two elements in $\mathcal{H}_1, \mathcal{H}_2$. For any $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$ we define the linear operator $g \otimes f \in \mathcal{S}(\mathcal{H}_1, \mathcal{H}_2)$, pointwise by

$$g \otimes f[h] := g\langle f, h\rangle_1 \qquad \forall h \in \mathcal{H}_1. \tag{2.2}$$

By virtue of this definition it is possible to endow $\mathcal{S}(\mathcal{H}_1, \mathcal{H}_2)$ with a particularly natural basis: If $\{f_n\}_{n\in\mathbb{N}}, \{g_n\}_{n\in\mathbb{N}}$ are ONBs of $\mathcal{H}_1$, $\mathcal{H}_2$ respectively, then the set $\{g_n \otimes f_m\}_{n,m\in\mathbb{N}}$ is an ONB of $\mathcal{S}(\mathcal{H}_1, \mathcal{H}_2)$. Finally, we notice that the outer product notation can be used to restate the spectral theorem for a compact, symmetric operator L in (2.1) as follows

$$L = \sum_{n\in\mathbb{N}} \eta_n f_n \otimes f_n.$$

In the next step we bring to bear these notations to the analysis of the functional regression problem (1.1).

2.2 The functional linear model

In this Section we introduce the functional, linear regression model (1.1) in a more rigorous way. Let $\mathcal{T} \subset \mathbb{R}^d$ denote a compact and non-empty set and μ_1, μ_2 measures defined on some σ-algebra on $\mathcal{T}$. Furthermore define $H_1 := (L^2(\mathcal{T}), \mu_1)$ and $H_2 := (L^2(\mathcal{T}), \mu_2)$ as the spaces of all measurable, real-valued functions on $\mathcal{T}$, that are square integrable w.r.t. μ_1 and μ_2, respectively. Equipped with the inner products

$$\langle f, g\rangle := \int_{\mathcal{T}} f(t)g(t)d\mu_i(t) \qquad f, g \in H_i \quad (i = 1, 2)$$

H_1 and H_2 are Hilbert spaces. Notice that the inner product $\langle f, g\rangle$ depends on the index $i = 1, 2$, but for the sake of simplicity we do not make this explicit. Accordingly, the norms induced by the inner products are denoted by $\|\cdot\|$.

This general setup includes many of the standard scenarios treated in the related literature. For instance to retrieve the model (1.2) with functional regressors and scalar responses (see Hall and Horowitz, 2007) it suffices to set $\mathcal{T} = [0, 1]$, $\mu_1 = \lambda$ (the Lebesgue measure) and $\mu_2 = \delta_1$ (the Dirac measure at the point

1). Another typical setting is to choose both measures as the Lebesgue measure, which gives functional inputs and outputs (see, for example, Yao et al., 2005, among many others). Further important non-standard cases such as spatio-temporal functions with continuous time and discrete space components (see Constantinou et al., 2017) can be accommodated as well.

Let $(X_1, Y_1), \ldots, (X_N, Y_N)$ denote N observations from a time series $\{(X_n, Y_n)\}_{n\in\mathbb{Z}} \subset H_1 \times H_2$, which are generated according to the linear model (1.1), that is

$$Y_n = SX_n + \varepsilon_n \qquad n = 1, ..., N, \tag{2.3}$$

where $S \in \mathcal{S}(H_1, H_2)$ denotes the (unknown) slope parameter and $\varepsilon_n \in H_2$ an observational error. By virtue of the outer product (see Section 2.1) it is possible to transform this linear model into a version, which is more suitable to inference about the slope parameter. More precisely, "multiplying" (2.3) by X_n from the right gives

$$Y_n \otimes X_n = SX_n \otimes X_n + \varepsilon_n \otimes X_n \qquad n = 1, ..., N. \tag{2.4}$$

Under the assumption $\mathbb{E}\|X_n\|^2, \mathbb{E}\|\varepsilon_n\|^2 < \infty$ the operators $SX_n \otimes X_n, \varepsilon_n \otimes X_n$ are random elements in $\mathcal{S}(H_1, H_2)$. Moreover, if the random functions X_n, ε_n are also centered, taking expectations on both sides of (2.4) gives

$$\mathbb{E}Y_n \otimes X_n = S\Gamma + \mathbb{E}\varepsilon_n \otimes X_n. \tag{2.5}$$

Here $\Gamma := \mathbb{E}X_n \otimes X_n$ is the covariance operator of X_n (recall that the sequence $\{X_n\}_{n\in\mathbb{N}}$ of regressors is stationary). Note that we merely assume centered regressors for ease of presentation and adaption to the non-centered case is simple (for details see Remark 3.8). Under the additional assumption of weak exogeneity, i.e., $\mathbb{E}\varepsilon_n \otimes X_n = 0$, equation (2.5) entails the fundamental identity

$$\mathbb{E}Y_n \otimes X_n = S\Gamma. \tag{2.6}$$

The task of recovering the operator S from equation (2.6) is non-trivial, even if we knew the "true" expectation $S\Gamma = \mathbb{E}Y_n \otimes X_n$ and the covariance operator Γ. One obvious condition for a complete recovery of S is identifiability, which is satisfied, if Γ is an injective operator. However, even in this case, as Γ is compact, its inverse must be unbounded and hence can only be defined on a dense linear subspace. We refer the interested reader to Dunford and Schwartz (1958) and Weidmann (1980) for a detailed discussion of (un)bounded operators.

A remedy for this problem is given by the application of a regularized inverse, i.e. a sequence of continuous operators $\{\Gamma_k^\dagger\}_{k\in\mathbb{N}}$, converging pointwise to Γ^{-1}. Of course this means that $|||\Gamma_k^\dagger|||_{\mathcal{L}} \to \infty = |||\Gamma^{-1}|||_{\mathcal{L}}$, but for each finite k the operator $S\Gamma\Gamma_k^\dagger$ is well defined on the whole space. Moreover, for sufficiently large k we expect that $S \approx S\Gamma\Gamma_k^\dagger$ in the sense that $|||S - S\Gamma\Gamma_k^\dagger|||$ becomes arbitrarily small. Let

$$\Gamma := \sum_{i=1}^{\infty} \lambda_i e_i \otimes e_i,$$

denote the spectral decomposition of the operator Γ, with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... > 0$ and corresponding eigenfunctions $e_1, e_2, ...$ A typical example of a regularized inverse operator is given by the spectral cut-off

regularizer

$$\Gamma_k^\dagger := \sum_{i=1}^{k} \frac{1}{\lambda_i} e_i \otimes e_i,$$

which evidently has operator norm $|||\Gamma_k^\dagger|||_{\mathcal{L}} = \lambda_k^{-1} < \infty$. We also point out that $\Gamma\Gamma_k^\dagger =: \Pi_k$, where Π_k is the projection on the space spanned by the first k eigenfunctions $e_1, ..., e_k$ of Γ. Notice that, if this was the whole problem, we could simply choose a large, but finite k and receive an arbitrarily precise approximation of S via $S\Gamma\Gamma_k^\dagger$. However, in practice neither the true expectation $\mathbb{E}Y_n \otimes X_n$, nor the true covariance operator Γ are known and have to be estimated from the data. For this purpose we define

$$\hat{\Gamma}_N := \frac{1}{N} \sum_{n=1}^{N} X_n \otimes X_n \tag{2.7}$$

as the standard estimate of the covariance operator Γ and $\frac{1}{N}\sum_{n=1}^{N} Y_n \otimes X_n$ as estimate of $\mathbb{E}Y_n \otimes X_n$. This gives an empirical analogue of equation (2.6), that is

$$\frac{1}{N} \sum_{n=1}^{N} Y_n \otimes X_n = S\hat{\Gamma}_N + U_N, \tag{2.8}$$

where

$$U_N := \frac{1}{N} \sum_{n=1}^{N} \varepsilon_n \otimes X_n \tag{2.9}$$

is a remainder term, arising from (2.4). Note that the identity (2.8) provides a way of estimating $S\Gamma$. We define the empirical version of the regularized inverse by

$$\hat{\Gamma}_k^\dagger := \sum_{i=1}^{k} \frac{1}{\hat{\lambda}_i} \hat{e}_i \otimes \hat{e}_i, \tag{2.10}$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq 0$ are the ordered eigenvalues of $\hat{\Gamma}_N$ and $\hat{e}_1, \hat{e}_2, ...$ the corresponding eigenfunctions. An estimator of the operator S is now given by

$$\hat{S}_N := \frac{1}{N} \sum_{n=1}^{N} Y_n \otimes X_n \hat{\Gamma}_k^\dagger = S\hat{\Gamma}_N \hat{\Gamma}_k^\dagger + U_N \hat{\Gamma}_k^\dagger = S\hat{\Pi}_k + U_N \hat{\Gamma}_k^\dagger \ , \tag{2.11}$$

where $\hat{\Pi}_k$ is the projection on the subspace spanned by the the first k eigenfunctions of the empirical covariance operator $\hat{\Gamma}_N$. This equation differs notably from the ideal $S\Gamma\Gamma_k^\dagger = S\Pi_k$ by the noise term $U_N\hat{\Gamma}_k^\dagger$ (which makes it a statistical inverse problem; see the discussion in the introduction). If k is large compared to N this remainder can potentially spoil the estimate, because the noise U_N is amplified by the regularized inverse $\hat{\Gamma}_k^\dagger$. Consequently the solution of the inverse problem as described in model (2.4), features a trade-off between regularization parameter k and sample size N.

As a corollary of our later discussion we will get a consistency result for $\hat{S}_N$ under suitable regularity conditions. For works specifically aimed at reconstructing the operator S see, for instance, Hall and Horowitz (2007), Benatia et al. (2017) and Imaizumi and Kato (2018).

3 Statistical inference for the location of S

In this section we introduce the concept of relevant hypotheses for the location of S and discuss the assumptions that are made throughout this paper. Furthermore we revisit the problems in deriving a weak convergence result for the estimator $\hat{S}_N$ as described in Cardot et al. (2007), Crambes and Mas (2013) and suggest a new technique - the smoothness shift - to grapple with them. Based on this idea, we establish an invariance principle for the estimated distance $|||\hat{S}_N - S_0|||^2$, which is used to develop pivotal statistics for testing relevant hypotheses.

3.1 Relevant differences in the slope

A typical concern in the context of model (1.1) is the comparison of the true slope S with some hypothesized operator $S_0 \in \mathcal{S}(H_1, H_2)$. This problem is often addressed by constructing statistical tests for the hypotheses

$$H_0 : S = S_0 \quad \text{versus} \quad H_1 : S \neq S_0 \ . \tag{3.1}$$

These hypotheses may for instance be used with $S_0 = 0$, to determine the explanatory power of the model, or with a slope S_0 from a theoretical model. Various tests have been devised for these (or related) hypotheses, such as in Cardot et al. (2003, 2004), where the cross covariance operator $\mathbb{E}Y_1 \otimes X_1 = S\Gamma$ is used to test the mathematically equivalent null hypothesis $S\Gamma = S_0\Gamma$. In a similar spirit Hilgert et al. (2013) propose minimax optimal adaptive tests based on projections of Y onto the principal components of Γ or Kong et al. (2016) employ traditional tests (such as the F-test) on finite dimensional subspaces, to validate model fit.

Although from a decision theoretical perspective all of these methods define consistent tests for the hypotheses (3.1), they have the drawback of telling us little about the actual proximity of the operators S and S_0. For example a test for H_0, based on the quantity $|||S\Gamma - S_0\Gamma|||$ is difficult to interpret, as $|||S\Gamma - S_0\Gamma|||$ may be arbitrarily small, while in fact the true difference $|||S - S_0|||$ is arbitrarily large. In particular if a user decides to perform data analysis under the assumption $S = S_0$ after a test has not rejected the hypothesis $S\Gamma = S_0\Gamma$, there is no guarantee that S_0 is indeed a good approximation of S. This insight has motivated some of the contemporary approaches to confidence regions for S (see the discussion in the introduction), where even a slower than parametric convergence rate is accepted, in return for an inference method, based on the original slope operator S.

In this paper we take up this insight and base statistical inference directly on the measure $|||S - S_0|||$. Evidently the point hypothesis in (3.1) is equivalent to $|||S - S_0||| = 0$. However in this work, we want to investigate the "relevant hypotheses", given by

$$H_0^{\Delta} : |||S - S_0|||^2 \leq \Delta \quad \text{versus} \quad H_1^{\Delta} : |||S - S_0|||^2 > \Delta, \tag{3.2}$$

and

$$H_0^{\tilde{\Delta}} : |||S - S_0|||^2 \geq \tilde{\Delta} \quad \text{versus} \quad H_1^{\tilde{\Delta}} : |||S - S_0|||^2 < \tilde{\Delta}, \tag{3.3}$$

where $\Delta > 0$ and $\tilde{\Delta} > 0$ are predetermined thresholds. Our suggestion to replace the "classical" hypotheses in (3.1) by hypotheses of the form (3.2) or (3.3) has theoretical as well as practical reasons.

From a theoretical perspective, testing exact equality of S and S_0 (both of which are infinite dimensional objects) might be questionable, because it is rarely believed that the hypothesized slope coincides perfectly with the true one. Therefore, testing H_0 means testing a hypothesis, which is essentially known to be false. This point is important, because any consistent test will detect any arbitrarily small deviation from H_0 if the sample size is sufficiently large (see Berkson, 1938) and thus we expect any consistent test for H_0 to eventually reject the hypothesis. This problem is evaded by the consideration of relevant hypotheses (3.2), which only refer to sufficient proximity of S and S_0.

We also believe that the relevant hypotheses are more congruent with common interests of users, who are less concerned with perfect equality than with the practical issue of comparable performance. Often users are willing to trade - at least to some extend - statistical precision for a simpler model. In this sense the thresholds $\Delta, \tilde{\Delta}$ in the relevant hypotheses can be understood as the largest deviation between S and S_0, which is still acceptable for the user. This also highlights that the choice of the threshold will depend on the application in hand and is not an a priori question. Finally we point out that a formulation of the hypotheses in the form (3.3) might be preferred if one is interested to work under the assumption $S = S_0$. If the null hypothesis $H_0^{\tilde{\Delta}}$ is rejected at level α, the risk of erroneously assuming $|||S - S_0|||^2 \leq \tilde{\Delta}$, is controlled, which is not possible using the "classical" hypotheses in (3.1), because there is no symmetry in the problem.

Although the hypotheses (3.2) and (3.3) are different with respect to their statistical interpretation it will become clear later that from a mathematical point of view they are in some sense equivalent. Therefore, and also for the sake of brevity, we restrict ourselves to the development of testing procedures for the hypotheses in (3.2) and denote the null hypothesis as "no relevant deviation from S_0".

3.2 Assumptions

The theoretical results of this paper require several assumptions, which are explained and illustrated in this section.

Recall that a stationary sequence $\{Z_j\}_{j\in\mathbb{Z}}$ of random variables is called ϕ-mixing, if $\lim_{k\to\infty} \phi(k) = 0$, where

$$\phi(k) := \sup_{h\in\mathbb{Z}} \sup \big\{ |\mathbb{P}(F|E) - \mathbb{P}(F)| \,:\, E \in \sigma(Z_1, \ldots, Z_h), F \in \sigma(Z_{h+k}, Z_{k+h+1}, \ldots),\ \mathbb{P}(E) > 0 \big\}.$$

denotes the ϕ-dependence coefficients and $\sigma(Z_h, ..., Z_k)$, is the σ-algebra generated by $Z_h, \ldots, Z_k$ (see for instance Dehling et al., 2002).

Assumption 3.1.

(1) *Smoothness:* For some $\beta \geq 0$ the operators S and S_0 are elements of the smoothness class

$$\mathcal{C}(\beta, \Gamma) := \big\{ R\Gamma^\beta : R \in \mathcal{S}_2(H_1, H_2) \big\}.$$

(2) *Moments:* There exists some $\kappa > 0$, such that $\mathbb{E}\|X_1\|^{4+\kappa}$, $\mathbb{E}\|\varepsilon_1\|^{4+\kappa} < \infty$.

(3) *Dependence:* The sequence of random functions $\{(X_n, \varepsilon_n)\}_{n\in\mathbb{Z}}$ is centered, strictly stationary and ϕ-mixing, such that

$$\phi(1) < 1 \quad \text{and} \quad \sum_{h\geq 1} \sqrt{\phi(h)} < \infty$$

(4) *Coefficients*: There exists a finite constant $C > 0$, s.t. the inequality $\mathbb{E}|\langle X_1, e_j\rangle|^4 \leq C(\mathbb{E}|\langle X_1, e_j\rangle|^2)^2$ holds for any $j \in \mathbb{N}$.
(5) *Weak exogeneity:* $\mathbb{E}\varepsilon_1 \otimes X_1 = 0$.
(6) *Decay of eigenvalues and eigengaps:* For some $\gamma > 0$ and large enough $C > 0$, the eigenvalues of the covariance operator Γ satisfy

$$\lambda_k \leq Ck^{-\gamma} \qquad \text{and} \qquad \lambda_k - \lambda_{k+1} \geq C^{-1}k^{-\gamma-1} \quad \forall k \in \mathbb{N}.$$

(7) *Rates of regularization:* The regularization parameter $k = k(N)$ is chosen such that for some $\delta > 0$

$$\frac{k^{\gamma+1+\delta}}{\sqrt{N}} \to 0 \qquad \text{and} \qquad \frac{k^{\gamma\beta}}{\sqrt{N}} \to \infty.$$

Remark 3.2.
(a) Assumption (1) is a smoothness condition on the slope operators S, S_0, w.r.t. the principal components of Γ. To see this let $S = R\Gamma^\beta$ and $x \in H_1$. It follows that $Sx = R\Gamma^\beta x = Ry$, where $y = \sum_{r\in\mathbb{N}} e_r(\lambda_r^\beta \langle x, e_r\rangle)$. Evidently the L^2-coefficients $(\lambda_r^\beta\langle x, e_r\rangle)$ of y decay faster than those of x, as they are weighted by a power of the decaying eigenvalues. In this sense y is smoother than x and a larger value of β translates into lighter coefficients and thus more smoothing. In this way Sx can be understood as the application of an integral operator R to a smoothed version of x. Assumption 3.1(1) was also considered in Benatia et al. (2017) in their study of the Tikhonov regularization, where it was denoted by the common label of *source condition*. At the beginning of their Section 3 the smoothing effect of Γ^β is explored by various examples. In the following calculations we demonstrate that Assumption 3.1(1) can be translated into *fast decaying tails* of the operator S ", which is another standard way of stating smoothness in the literature. Consider the application of R to a basis function e_q of Γ

$$Re_q = R\Gamma^\beta\Gamma^{-\beta}e_q = S\Gamma^{-\beta}e_q.$$

Notice that $\Gamma^{-\beta}e_q = \lambda_q^{-\beta}e_q$ is indeed well defined. We can now express Re_q as

$$Re_q = S\Gamma^{-\beta}e_q = \Big[\sum_{i,j\in\mathbb{N}} s_{i,j}e_i \otimes e_j\Big]\Big[\sum_{k\in\mathbb{N}} \lambda_k^{-\beta}e_k \otimes e_k\Big]e_q = \sum_{i\in\mathbb{N}} s_{i,q}\lambda_q^{-\beta}e_i,$$

where $s_{i,j} := \langle S, e_i \otimes e_j\rangle$ (with the inner product on the space of Hilbert–Schmidt operators, see Section 2.1). Now the squared Hilbert–Schmidt norm of R equals

$$\begin{aligned}\infty > |||R|||^2 &= \sum_{q\in\mathbb{N}}\langle Re_q, Re_q\rangle = \sum_{q\in\mathbb{N}}\langle\sum_{i\in\mathbb{N}} s_{i,q}\lambda_q^{-\beta}e_i, \sum_{l\in\mathbb{N}} s_{l,q}\lambda_q^{-\beta}e_l\rangle = \sum_{q,l\in\mathbb{N}} \lambda_q^{-2\beta}s_{l,q}^2 \\ &= \sum_{q,l\in\mathbb{N}} \lambda_q^{-2\beta}\langle Se_q, e_l\rangle^2 = \sum_{q\in\mathbb{N}} \lambda_q^{-2\beta}\|Se_q\|^2, \end{aligned} \tag{3.4}$$

where we have used Parseval's identity in the last step. In the scalar response model (1.2) one has $\|Se_q\|^2 = \langle\varphi_S, e_q\rangle^2$. Thus the summability in (3.4) is a smoothness condition for φ_S. In this form it has been used by Hall and Horowitz (2007) (see equation (3.3) in that paper). In the more general model (1.1) the decay of $\|Se_q\|^2$ was considered as a smoothness condition in Crambes and Mas (2013) (see their Definition 3).

(b) Assumptions 3.1(2) - (5) are required to derive a weak convergence result stated in Theorem 3.5. The existence of moments of larger order than 4 is typical for proving second order, weak invariance principles (it corresponds to the assumption of more than second moments for the first order; see Berkes et al., 2013). The mixing assumption is weaker than those in the related literature, where almost exclusively i.i.d. observations are considered, (see Hall and Horowitz, 2007; Crambes and Mas, 2013; Benatia et al., 2017; Imaizumi and Kato, 2019; Babii, 2020, among others). Assumption 3.1(4) regarding the moments of the coefficients $\langle X_1, e_j \rangle$ is standard in the literature (see for example Hall and Horowitz, 2007; Crambes and Mas, 2013; Imaizumi and Kato, 2019) and is needed for technical reasons. We use it in the proof of Lemma A.3, part ii). Assumption 3.1(5) regarding the exogeneity is again weaker than in most of the literature. Here often strong exogeneity is required (see the literature cited before), where the work of Benatia et al. (2017) constitutes an important exception.

(c) Assumption 3.1(6) guarantees a polynomial decay rate for the eigenvalues of Γ, that is $\lambda_k \sim k^{-\gamma}$. More important than the precise rate of decay is the assumption on the eigengaps, which have to be controlled for identifiability reasons. Assumptions of this type are standard in the literature, in particular in the analysis of spectral cut-off estimators (see, Hall and Horowitz, 2007; Qiao et al., 2019, among others), even though they are sometimes made implicitly (see Lemma 12 in Crambes and Mas, 2013).
The two decay rates in Assumption 3.1(7) expose the trade-off inherent in the choice of k. On the one hand k has to increase slowly enough, such that the k-th eigenvalue λ_k can be distinguished from λ_{k+1}. This means that the k-th eigengap of size $k^{-\gamma-1}$ is of larger order than the estimation error of size $1/\sqrt{N}$. Our assumption is almost sharp in the sense that we assume $k^{\gamma+1}/\sqrt{N}$ to decay at some arbitrarily slow polynomial rate in N. We use this additional leverage to derive not only a CLT but a stronger weak invariance principle, where remainders have to be controlled uniformly; see Lemma A.3. A sharp version has been used for confidence bands in the scalar response model by Imaizumi and Kato (2019). On the other hand, k has to increase fast enough, such that the asymptotic bias is negligible, more precisely

$$|||S - S\Pi_k||| = |||R\Gamma^\beta[Id - \Pi_k]||| \leq |||R|||\,|||\Gamma^\beta[Id - \Pi_k]|||_{\mathcal{L}} = |||R|||\lambda_{k+1}^\beta = \mathcal{O}(k^{-\gamma\beta}) = o(1/\sqrt{N}).$$

It can be shown that the above bound is sharp for general operators and hence the bias rate cannot be improved upon. Notice that the two Assumptions on k can be simultaneously fulfilled if and only if

$$\beta > 1 + 1/\gamma.$$

3.3 Main results

In order to develop a statistical test for the relevant hypotheses defined in (3.2) it is reasonable to estimate the difference $|||S - S_0|||^2$. A natural estimator is given by $|||\hat{S}_N - S_0\hat{\Pi}_k|||^2$. While it is also possible to replace $S_0\hat{\Pi}_k$ by S_0 in the subsequent theory, we prefer to work with $S_0\hat{\Pi}_k$ as it does not seem sensible to compare S_0 along dimensions to S, where no estimate for S exists (this common sense approach is also supported by simulations). In order to define a consistent and (asymptotic) level-α test, we are interested

in the weak convergence of the difference

$$\sqrt{N}\big(|||\hat{S}_N - S_0\hat{\Pi}_k|||^2 - |||S - S_0|||^2\big). \tag{3.5}$$

The standard approach to this problem would be to: first establish weak convergence of the difference $\sqrt{N}(\hat{S}_N - S)$ in the space $\mathcal{S}(H_1, H_2)$; then deduce weak convergence of the test statistic in (3.5) by applying the Delta method (see Section 3.9 in van der Vaart and Wellner (1996)) to the mapping $S \to |||S - S_0|||^2$. Notice that, using the OLS estimator, this method works for finite dimensional linear regression. However this approach fails in the context of functional regression problems, as it is not possible to find a standardizing sequence, say $\{a_N\}_{N\in\mathbb{N}}$, such that the difference $a_N(\hat{S}_N - S)$ converges weakly to a non-degenerate limit, if k converges to infinity with the sample size, which is necessary to obtain an asymptotically vanishing bias (see, for example, Crambes and Mas, 2013). More precisely, if k is fixed one can prove that $\sqrt{N}(\hat{S}_N - S)\Pi_k$ converges weakly to a Gaussian random vector. A similar result was derived by Benatia et al. (2017) for a different regularization method. However these authors likewise concluded that for decaying regularization, i.e. $k \to \infty$ as $N \to \infty$ the sequence $\sqrt{N}(\hat{S}_N - S)$ has a degenerate limit caused by an inflation of the error variance.
Nevertheless the fact that no weak convergence of $\sqrt{N}(\hat{S}_N - S)$ in the space $\mathcal{S}(H_1, H_2)$ can be established does not necessarily imply that the difference in (3.5) cannot converge weakly. Indeed we will demonstrate that the mapping $S \to |||S - S_0|||^2$ has a smoothing effect on $\hat{S}_N$. Therefore the inflation of the observation error U_N (defined in (2.9)) is compensated and it is possible to establish weak convergence of (3.5) with a normally distributed limit. The precise statement will be given in Proposition 3.3 below. To get an intuition how this smoothing works note that by the third binomial formula in Hilbert spaces we have

$$\sqrt{N}\big(|||\hat{S}_N - S_0\hat{\Pi}_k|||^2 - |||(S - S_0)\Pi_k|||^2\big) = \langle\sqrt{N}[\hat{S}_N - S_0\hat{\Pi}_k - (S - S_0)\Pi_k], \hat{S}_N - S_0\hat{\Pi}_k + (S - S_0)\Pi_k\rangle.$$

After some careful bounding of the error terms (recall that the left side of the inner product asymptotically degenerates), we can show that this equals

$$2\sqrt{N}\langle\hat{S}_N - S_0\hat{\Pi}_k - (S - S_0)\Pi_k, S - S_0\rangle + o_{\mathbb{P}}(1).$$

By Assumption 3.1(1) there exist operators $R, R_0 \in \mathcal{S}(H_1, H_2)$, such that $S = R\Gamma^\beta, S_0 = R_0\Gamma^\beta$. Hence we can perform the following *smoothness shift*, moving smoothness in the form of Γ^β from the second to the first component of the inner product, i.e.

$$\begin{aligned}2\sqrt{N}\langle\hat{S}_N - S_0\hat{\Pi}_k - (S - S_0)\Pi_k, S - S_0\rangle =& 2\sqrt{N}\langle\hat{S}_N - S_0\hat{\Pi}_k - (S - S_0)\Pi_k, [R - R_0]\Gamma^\beta\rangle \\ =& 2\sqrt{N}\langle[\hat{S}_N - S_0\hat{\Pi}_k - (S - S_0)\Pi_k]\Gamma^\beta, R - R_0\rangle.\end{aligned}$$

It turns out that the smoothing effect of Γ^β on the left stops the error inflation and thus weak convergence to a non-degenerate and (with some technical linearization) normally distributed limit can be proved. Intuitively the smoothing works, because

$$\hat{S}_N\Gamma^\beta = \frac{1}{N}\sum_{n=1}^{N} Y_n \otimes X_n\hat{\Gamma}_k^\dagger\Gamma^\beta \approx \frac{1}{N}\sum_{n=1}^{N} Y_n \otimes X_n\Gamma^{\beta-1},$$

i.e. the regularized inverse $\hat{\Gamma}_k^\dagger$ and the shifted operator Γ^β "cancel out" to $\Gamma^{\beta-1}$, thus eliminating the pathology of the asymptotically unbounded operator $\hat{\Gamma}_k^\dagger$. If the term on the right is centered and standardized by $\sqrt{N}$ it is asymptotically normal. The price we pay for this non-standard approach is a more elaborate proof, where many difficult remainders have to be controlled. As announced we now formulate the precise result.

Proposition 3.3. *Under the Assumptions presented in Section 3.2, it holds that*

$$T_N = \sqrt{N}\left(|||\hat{S}_N - S_0\hat{\Pi}_k|||^2 - |||(S-S_0)\Pi_k|||^2\right) \xrightarrow{d} \mathcal{N}(0,\tau^2), \tag{3.6}$$

where the long-run variance τ^2 is defined by

$$\begin{aligned}\tau^2 :=& 4\Bigg\{\sum_{h\in\mathbb{Z}} \mathbb{E}\Big[\langle (R-R_0)L[X_0\otimes X_0-\Gamma], R-R_0\rangle \langle (R-R_0)L[X_h\otimes X_h-\Gamma], R-R_0\rangle\Big] \\ &+ 2\mathbb{E}\Big[\langle (R-R_0)L[X_0\otimes X_0-\Gamma], R-R_0\rangle \langle \varepsilon_h\otimes X_h\Gamma^{\beta-1}, R-R_0\rangle\Big] \\ &+ \mathbb{E}\Big[\langle \varepsilon_0\otimes X_0\Gamma^{\beta-1}, R-R_0\rangle \langle \varepsilon_h\otimes X_h\Gamma^{\beta-1}, R-R_0\rangle\Big]\Bigg\}.\end{aligned} \tag{3.7}$$

Here the map L is given in Definition A.4 of the Appendix.

Using Proposition 3.3, we could in principle construct a test for the hypothesis of no relevant deviation, presented in (3.2), by rejecting the null hypothesis, whenever

$$\sqrt{N}\big(|||\hat{S}_N - S_0\hat{\Pi}_k|||^2 - \Delta\big) > \tau\Phi^{-1}(1-\alpha), \tag{3.8}$$

where Φ^{-1} is the quantile function of a standard normal distribution and $\alpha\in(0,1)$ denotes the nominal level. This decision yields indeed a test which is asymptotically consistent and keeps its nominal level asymptotically. To see this we use the expansion

$$\sqrt{N}\big(|||\hat{S}_N - S_0\hat{\Pi}_k|||^2 - \Delta\big) = T_{1N} + T_{2N} + T_{3N}\ , \tag{3.9}$$

where

$$\begin{aligned}T_{1N} &= \sqrt{N}\big(|||\hat{S}_N - S_0\hat{\Pi}_k|||^2 - |||(S-S_0)\Pi_k|||^2\big),\\ T_{2N} &= \sqrt{N}\big(|||(S-S_0)\Pi_k|||^2 - |||S-S_0|||^2\big),\\ T_{3N} &= \sqrt{N}\big(|||S-S_0|||^2 - \Delta\big).\end{aligned}$$

By Proposition 3.3 the first term T_{1N} in (3.9) converges weakly to a centered normal distribution with variance τ^2. The term T_{2N} is the bias and asymptotically vanishes (see discussion of Assumption 3.1(7)). The third term T_{3N} is also deterministic. In the *interior* of the null hypothesis, that is $|||S-S_0|||^2 < \Delta$, it converges to $-\infty$ and thus asymptotically no rejection occurs for $N\to\infty$. On the *boundary* of the hypothesis, that is $|||S-S_0|||^2 = \Delta$, it vanishes and we get $\sqrt{N}\big(|||\hat{S}_N - S_0\hat{\Pi}_k|||^2 - \Delta\big) = T_{1N} + o_{\mathbb{P}}(1) \to \mathcal{N}(0,\tau^2)$. Consequently, the test (3.8) has asymptotic level α in this case. Notice that the bias T_{2N} is always non-positive which means that small choices of k (resulting in larger bias) invariably make the test

more conservative. Finally, under the alternative the term T_{3N} diverges to ∞ and thus rejection occurs with probability converging to 1 (asymptotic consistency). In the following remark we briefly explain how the decomposition (3.9) can be used for a more refined analysis with respect to local alternatives.

Remark 3.4. Consider the local alternative of $S = \tilde{S} + cH/\sqrt{N}$, where both $H, \tilde{S} \in \mathcal{C}(\beta, \Gamma)$ (see Assumption 3.1(1)) are operators, $c > 0$ is a scaling factor and $\Delta = |||\tilde{S} - S_0|||^2$ demarcates the boundary of the hypothesis. Furthermore assume that $\langle \tilde{S} - S_0, H \rangle =: \Lambda > 0$. The last requirement is necessary, such that we are indeed under the alternative ($\Lambda < 0$ corresponds to the hypothesis) and thus

$$|||S - S_0|||^2 = |||\tilde{S} - S_0|||^2 + 2c\langle H, \tilde{S} - S_0\rangle/\sqrt{N} + c^2|||H|||^2/N = \Delta + 2c\Lambda/\sqrt{N} + \mathcal{O}(1/N) > \Delta.$$

Now suppose that the Assumptions of Proposition 3.3 hold. We apply the test, defined in (3.8) in this situation and let $p_{H_1^\Delta}(c)$ denote the probability of rejection. It then follows that

$$\lim_{N\to\infty} p_{H_1^\Delta}(c) > \alpha \quad \Leftrightarrow \quad c > 0 \qquad \text{and} \qquad \lim_{c\to\infty}\lim_{N\to\infty} p_{H_1^\Delta}(c) = 1.$$

Both results follow from the decomposition (3.9). It is not difficult to show that T_{1N} converges to the same normal distribution as in the case of $S = \tilde{S}$, that T_{2N} asymptotically vanishes and that $T_{3N} = 2c\Lambda + o(1) > 0$ is non-vanishing and (asymptotically) scales linearly with c. Consequently the test in (3.8) is able to detect local alternatives converging to the null hypothesis at a rate of $1/\sqrt{N}$. We also point out that all subsequently presented, self-normalized tests directly inherit this property, as the numerator of the normalized statistic can be decomposed as in (3.9).

Note that the test (3.8) provides an attractive decision rule for the hypothesis (3.2) supposing that a reliable estimate of the variance τ^2 is available. Unfortunately, even in the case of independent observations this quantity is painfully complex to estimate. It requires not only estimation of Γ, but also of the fourth order structure of regressors and errors, a linearization map L depending on all eigenvalues and eigenfunctions of the operator Γ (an object which depends inversely on the small eigengaps) as well as knowledge about the operators R and R_0 in Assumption 3.1(1). What is difficult for i.i.d. data is almost infeasible in the case of dependent data. In this case τ^2 is a long-run variance, which requires besides the estimation of all the mentioned entities the determination of a bandwidth, capturing the sequential dependence of the regressors and errors.

Given the impracticality and instability of such an estimate we pursue the different approach of self-normalization in the following section. The technical prerequisite for this procedure is the derivation of a weak invariance principle, generalizing Proposition 3.3. For this purpose we introduce a sequential version of the statistic $\hat{S}_N$ which is defined similarly as the original, with the difference that - instead of all N observations - only the observations $(X_1, Y_1), \ldots, (X_{\lfloor \xi N \rfloor}, Y_{\lfloor \xi N \rfloor})$ for $\xi \in (0, 1]$ are used for estimation. To be precise we define the sequential covariance estimator

$$\hat{\Gamma}_N[\xi] = \frac{1}{N}\sum_{n=1}^{\lfloor \xi N \rfloor} X_n \otimes X_n. \tag{3.10}$$

Furthermore, we define the sequential estimators of the eigenvalues and eigenfunctions of Γ, denoted by $\hat{\lambda}_i[\xi], \hat{e}_i[\xi]$ as the eigenvalues and eigenfunctions of the operator $\hat{\Gamma}_N[\xi]$ (where the eigenvalues are again

assumed to be in non-increasing order). With these estimators we set

$$\hat{\Gamma}_k^\dagger[\xi] = \sum_{i=1}^{k} \frac{1}{\hat{\lambda}_i[\xi]} \hat{e}_i[\xi] \otimes \hat{e}_i[\xi] \qquad \text{and} \qquad \hat{\Pi}_k[\xi] = \sum_{i=1}^{k} \hat{e}_i[\xi] \otimes \hat{e}_i[\xi]. \tag{3.11}$$

Finally the sequential estimator of S is given by

$$\hat{S}_N[\xi] := \frac{1}{N} \sum_{n=1}^{\lfloor \xi N \rfloor} Y_n \otimes X_n \hat{\Gamma}_k^\dagger[\xi]. \tag{3.12}$$

Note that in the case of $\xi = 1$ these estimators are identical to their non-sequential counterparts $\hat{\Gamma}_N$, $\hat{\Gamma}_k^\dagger$ and $\hat{S}_N$ defined in (2.7), (2.10) and (2.11), respectively, and that we do not adapt $k = k(N)$ to ξ. Throughout this paper we will use the notations $\hat{\Gamma}_N$ and $\hat{\Gamma}_N[1]$ simultaneously. We can now state the weak invariance principle generalizing Proposition 3.3.

Theorem 3.5. *Under our Assumptions 3.1, it holds for any compact interval $I \subset (0,1]$ that*

$$\left\{ \sqrt{N}\xi \left(|||\hat{S}_N[\xi] - S_0 \hat{\Pi}_k[\xi]|||^2 - |||(S - S_0)\Pi_k|||^2 \right) \right\}_{\xi \in I} \xrightarrow{d} \{\tau \mathbb{B}(\xi)\}_{\xi \in I},$$

as $N \to \infty$, where $\mathbb{B}$ is a standard Brownian motion and the long-run variance τ^2 is defined in (3.7).

3.4 A pivotal test statistic

In the last section we have derived a weak invariance principle for the estimated deviation measure in (3.5). While a central limit theorem is theoretically sufficient to construct a test for the hypothesis (3.2), as we have seen in the discussion of Proposition 3.3, the estimation of the long-run variance τ^2 is infeasible in applications. In this section we circumvent the problem of estimating τ^2, by a self-normalization approach, based on the weak invariance principle in Theorem 3.5. For this purpose, define for $0 < a < 1$ the interval $I = [a, 1]$, let ν be a probability measure on I and consider the normalizer

$$\hat{V}_N := \left\{ \int_I \xi^4 \left(|||\hat{S}_N[\xi] - S_0 \hat{\Pi}_k[\xi]|||^2 - |||\hat{S}_N - S_0 \hat{\Pi}_k|||^2 \right)^2 d\nu(\xi) \right\}^{1/2}. \tag{3.13}$$

The next corollary is a consequence of Theorem 3.5 and the continuous mapping Theorem. It can be viewed as a standardized version of Proposition 3.3.

Corollary 3.6. *Suppose that the assumptions of Theorem 3.5 hold and that $\tau > 0$. Then the weak convergence*

$$\frac{\left(|||\hat{S}_N - S_0 \hat{\Pi}_k|||^2 - |||(S - S_0)\Pi_k|||^2 \right)}{\hat{V}_N} \xrightarrow{d} W := \frac{\mathbb{B}(1)}{\left\{ \int_I \xi^2 (\mathbb{B}(\xi) - \xi \mathbb{B}(1))^2 d\nu(\xi) \right\}^{1/2}} \tag{3.14}$$

holds, where $\mathbb{B}$ is a standard Brownian motion on the interval $[0,1]$.

We point out that the quantiles of the distribution of W can be readily simulated using the Fourier representation of the Brownian motion. A typical choice for the measure ν is a discrete uniform measure

on the set $\{1/T, 2/T, ..., (T-1)/T, 1\}$ for some $T \geq 2$. Simulations suggest that the choice of T has little impact on the statistical performance of the resulting procedure, while - of course - smaller values of T yield computational advantages (see Section 6).

In view of (3.8) and the subsequent discussion we now define a decision rule for the hypothesis in (3.2) rejecting the null hypothesis in (3.2), whenever

$$\hat{W}_N(\Delta) := \frac{\sqrt{N}\big(|||\hat{S}_N - S_0\hat{\Pi}_k|||^2 - \Delta\big)}{\hat{V}_N} > q_{1-\alpha}, \tag{3.15}$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of the distribution of W in (3.14). The next theorem shows the validity of this test decision.

Theorem 3.7. *Under the assumptions of Corollary 3.6 the decision presented in* (3.15) *yields an asymptotic level-α and consistent test for the hypothesis* (3.2).

Remark 3.8.

(1) In the theoretical results presented so far it is assumed that the regressors are centered, that is $\mathbb{E}X_1 = 0$. In reality it may well be that $\mathbb{E}X_1 = \mu \neq 0$ and therefore an empirical centering is necessary. More precisely we can introduce the sequential mean estimates (recall that $\xi \in [a, 1]$, $0 < a < 1$)

$$\bar{Y}_N[\xi] := \frac{1}{\lfloor N\xi \rfloor} \sum_{n=1}^{\lfloor N\xi \rfloor} Y_n, \quad \text{and} \quad \bar{X}_N[\xi] := \frac{1}{\lfloor N\xi \rfloor} \sum_{n=1}^{\lfloor N\xi \rfloor} X_n,$$

and consider the modified observations $Y_n - \bar{Y}_N[\xi]$ and $X_n - \bar{X}_N[\xi]$ in any of the sequential statistics introduced at the beginning of this section. It can be shown that all results presented so far remain correct in this case (we also employ this empirical centering in the simulation study in Section 6).

(2) It follows from the proof of Theorem 3.7 that $\lim_{N\to\infty} \mathbb{P}\big(\hat{W}_N(\Delta) > q_{1-\alpha}\big) = 0$ if $|||S - S_0|||^2 < \Delta$ (interior of the null hypothesis), while for $|||S - S_0|||^2 = \Delta$ (boundary of the null hypothesis) we have $\lim_{N\to\infty} \mathbb{P}\big(\hat{W}_N(\Delta) > q_{1-\alpha}\big) = \alpha$.

(3) It is easy to see that the test statistic $\hat{W}_N(\Delta)$ is a decreasing function of the threshold Δ. This means that rejection for some $\Delta > 0$ also entails rejection for all smaller thresholds and vice versa accepting the hypothesis for some threshold means acceptance for all larger values. Hence the interpretation for multiple values of Δ - if considered - is internally consistent.

(4) Similar results can also be obtained for other dependence concepts than $\phi-$mixing. For example, consider α-mixing processes (for a definition, see, for instance Dehling et al., 2002) and assume

$(3')$: The sequence $\{(X_n, \varepsilon_n)\}_{n\in\mathbb{Z}}$ is strictly stationary and α-mixing s.t. $\sum_{h\geq 1} h^{4/\kappa}\alpha(h) < \infty$

$(4')$: If c is the smallest even integer $c > 4 + \kappa$, then $\mathbb{E}|\langle X_1, e_j\rangle|^c \leq C(\mathbb{E}|\langle X_1, e_j\rangle|^2)^{c/2}$.

respectively, where $\alpha(h)$ denotes the α-mixing coefficient. Then all statements in this and the subsequent sections remain correct if the conditions (3) and (4) in Assumption 3.1 are replaced by $(3')$ and $(4')$, respectively. For technical details we refer the interested reader to Dehling et al. (2002) (covariance inequalities for $\alpha-$mixing in Hilbert spaces) and to Merlevède et al. (2006) (invariance principles under α-mixing).

4 Statistical inference for relevant prediction errors

In the previous section we have compared the slope operator S to a predetermined operator S_0, in terms of the Hilbert–Schmidt norm $|||S-S_0|||^2$. However, from a statistical perspective other deviation measures are at least equally important. One vital mode of comparison is, in how far the predictions of the two operators differ, which we discuss in this section. Prediction in finite and infinite dimensional linear models is a well-investigated subject. In the work most closely related to our own, Crambes and Mas (2013) considered the minimax prediction error of the spectral cut-off estimator $\hat{S}_N$, compared to the true slope S. The focus in our work is different, as we want to compare the predictive properties of the true slope S, with the hypothesized operator S_0. More specifically we are interested in the quantity $\mathbb{E}\|SX - S_0X\|^2$, where the expectation is taken with respect to a regressor X, distributed as X_1. A simple calculation, using the trace representation of inner products and its properties (see Section 13.5 in Horváth and Kokoszka (2012)) shows that

$$\mathbb{E}\|SX - S_0X\|^2 = |||S\Gamma^{1/2} - S_0\Gamma^{1/2}|||^2. \tag{4.1}$$

Therefore we are comparing smoothed versions of the slope operators. We point out that even though the inequality

$$|||S\Gamma^{1/2} - S_0\Gamma^{1/2}|||^2 \leq |||\Gamma^{1/2}|||_{\mathcal{L}}^2 |||S - S_0|||^2$$

implies that small differences between S and S_0 result in small prediction errors, the converse is not true. In particular small prediction errors may be found in operators, that vastly differ in the Hilbert–Schmidt norm.

We now formulate the hypothesis of no relevant prediction error as

$$^{\text{pred}}H_0^{\Delta} : \mathbb{E}\|SX - S_0X\|^2 \leq \Delta \qquad ^{\text{pred}}H_1^{\Delta} : \mathbb{E}\|SX - S_0X\|^2 > \Delta, \tag{4.2}$$

where X has the same distribution as X_1. Again $\Delta > 0$ is a user determined threshold, where a deviation of more than Δ is considered scientifically relevant. In order to test this hypothesis we recall the identity (4.1) which suggests the natural estimator $|||\hat{S}_N\hat{\Gamma}_N^{1/2} - S_0\hat{\Pi}_k\hat{\Gamma}_N^{1/2}|||^2$ for the prediction error. Recall that the projection $S_0\hat{\Pi}_k$ can be replaced by the operator S_0, but projecting seems more sensible, because otherwise S_0 is compared to S along axes, which are not estimated. Compared to the statistic discussed in Section 3, we expect that the multiplication with $\hat{\Gamma}_N^{1/2}$ leads to an even stronger smoothing effect, which indeed manifests itself in weaker assumptions on the regularization parameter.

Assumption 4.1. *Rates of regularization:* The regularization parameter k satisfies for some $\delta > 0$

$$\frac{k^{\gamma+1+\delta}}{\sqrt{N}} \to 0 \qquad \text{and} \qquad \frac{k^{\gamma(\beta+1/2)}}{\sqrt{N}} \to \infty.$$

If Assumption 4.1 holds, the bias of the prediction error vanishes asymptotically, as

$$\begin{aligned} |||[S - S\Pi_k]\Gamma^{1/2}||| &= |||R\Gamma^{\beta}[I - \Pi_k]\Gamma^{1/2}||| \leq |||R|||\,|||\Gamma^{\beta}[I - \Pi_k]\Gamma^{1/2}|||_{\mathcal{L}} \\ &= |||R|||\lambda_{k+1}^{\beta+1/2} = \mathcal{O}(k^{-\gamma(\beta+1/2)}) = o(1/\sqrt{N}). \end{aligned}$$

Notice that compared to Assumption 3.1(7), Assumption 4.1 translates into weaker smoothness requirements for the operators S and S_0. In fact it implies $\beta > 1/2 + 1/\gamma$ (instead of $\beta > 1 + 1/\gamma$, because

$S \in \mathcal{C}(\beta - 1/2, \Gamma)$ already entails $S\Gamma^{1/2} \in \mathcal{C}(\beta, \Gamma)$). In applications this effect is reflected by smaller values of k in the spectral cut-off estimator for prediction compared to reconstruction. Nevertheless the representation

$$\hat{S}_N \hat{\Gamma}_N^{1/2} = S\hat{\Pi}_k \hat{\Gamma}_N^{1/2} + U_N(\hat{\Gamma}_k^{\dagger})^{1/2},$$

suggests, that inference for the prediction error remains a genuinely inverse problem. In particular we still observe an amplification of the observation error U_N by the regularized inverse, but to a weaker extend than in the case of reconstruction.

Recalling the definition of the sequential estimators (3.10), (3.11) and (3.12) in Section 3.3 we obtain the following invariance principle.

Theorem 4.2. *Under the Assumptions 3.1(1)-(6) and Assumption 4.1, it holds for any compact interval $I \subset (0, 1]$, that*

$$\Big\{\sqrt{N}\xi \Big(|||\hat{S}_N[\xi]\hat{\Gamma}_N[\xi]^{1/2} - S_0\hat{\Pi}_k[\xi]\hat{\Gamma}_N[\xi]^{1/2}|||^2 - |||(S - S_0)\Pi_k\Gamma^{1/2}|||^2 \Big)\Big\}_{\xi \in I} \xrightarrow{d} \{\tau^{\mathrm{pred}}\mathbb{B}(\xi)\}_{\xi \in I},$$

where the long-run variance $(\tau^{\mathrm{pred}})^2$ is defined as follows

$$\begin{aligned}(\tau^{\mathrm{pred}})^2 :=& 4\Big\{ \sum_{h \in \mathbb{Z}} \mathbb{E}\Big[\langle (R - R_0)\tilde{L}[X_0 \otimes X_0 - \Gamma], R - R_0\rangle \langle (R - R_0)\tilde{L}[X_h \otimes X_h - \Gamma], R - R_0\rangle\Big] \\ &+ 2\mathbb{E}\Big[\langle (R - R_0)\tilde{L}[X_0 \otimes X_0 - \Gamma], R - R_0\rangle \langle \varepsilon_h \otimes X_h\Gamma^{\beta-1}, R - R_0\rangle\Big] \\ &+ \mathbb{E}\Big[\langle \varepsilon_0 \otimes X_0\Gamma^{\beta-1}, R - R_0\rangle \langle \varepsilon_h \otimes X_h\Gamma^{\beta-1}, R - R_0\rangle\Big]\Big\}. \end{aligned} \tag{4.3}$$

Here the map $\tilde{L}$ is given in Definition A.4.

Next we define the adapted denominator

$$\hat{V}_N^{\mathrm{pred}} := \Big\{ \int_I \xi^4 \Big(|||\hat{S}_N[\xi]\hat{\Gamma}_N[\xi]^{1/2} - S_0\hat{\Pi}_k[\xi]\hat{\Gamma}_N[\xi]^{1/2}|||^2 - |||\hat{S}_N\hat{\Gamma}_N^{1/2} - S_0\hat{\Pi}_k\hat{\Gamma}_N^{1/2}|||^2 \Big)^2 d\nu(\xi) \Big\}^{1/2}, \tag{4.4}$$

and propose to reject the null hypothesis in (4.2), if

$$\hat{W}_N^{\mathrm{pred}}(\Delta) := \frac{\sqrt{N}\big(|||\hat{S}_N\hat{\Gamma}_N^{1/2} - S_0\hat{\Pi}_k\hat{\Gamma}_N^{1/2}|||^2 - \Delta\big)}{\hat{V}_N^{\mathrm{pred}}} > q_{1-\alpha}. \tag{4.5}$$

Theorem 4.3. *Suppose that the Assumptions 3.1(1)-(6), Assumption 4.1 hold and that the long-run variance τ^{pred} is positive. Then the decision rule in* (4.5) *defines a consistent, asymptotic level-α test for the hypothesis in* (4.2).

We conclude this part by comparing the weak convergence result of this Section to those derived in Crambes and Mas (2013) for prediction.

Remark 4.4. Crambes and Mas (2013) proved a weak convergence result in the case of i.i.d. observations and somewhat different assumptions than those used in this section. In their Theorem 9 (which is a generalization of Theorem 4.2 in Cardot et al. (2007)), they showed that for a random variable X distributed as X_1 and independent of the sequence $\{(X_n, \varepsilon_n)\}_{n \in \mathbb{Z}}$, the weak convergence

$$\sqrt{N/k}(\hat{S}_N X - SX) \xrightarrow{d} G \tag{4.6}$$

holds, where G is a centered Gaussian process on H_2, with covariance operator $\mathbb{E}\varepsilon_1 \otimes \varepsilon_1$. Notice the standardization of $\sqrt{N/k}$ instead of $\sqrt{N}$, which corresponds to the standard deviation of $U_N \Gamma_k^\dagger X$. This term naturally occurs (as second term) in the decomposition

$$\sqrt{N/k}(\hat{S}_N - S\Pi_k)X = \sqrt{N/k}\left\{S(\hat{\Pi}_k - \Pi_k)X + U_N \hat{\Gamma}_k^\dagger X\right\}.$$

Importantly the first term here is asymptotically negligible, which is not the case in our smoothed statistic. Indeed, in the L^2-statistic, after the smoothness shift is performed, the amplifying effect of the regularized inverse is eliminated, which yields the convergence rate $1/\sqrt{N}$ for both terms instead of $\sqrt{k/N}$. In view of these technical differences we have developed a separate asymptotic theory for the proof of Theorem 4.3 tailored to the study of relevant hypotheses and could not use the result in (4.6).

5 Change point analysis and two sample tests

In the context of dependent time series, functional data analysis is usually employed to model successive observations of a system over an extended time period. In this context it is natural to consider the stability of the data, e.g., by searching for change points in the mean (see e.g. Berkes et al., 2009), Chapter 6 in Horváth and Kokoszka (2012), Aston and Kirch (2012a) or Dette et al. (2020)) or in the second order structure, i.e. covariance operators (Jarušková, 2013), principle components (Dette and Kutta, 2021) or other features (Aue et al., 2020). For the linear regression model (1.1) stability concerns first and foremost the slope operator S. This problem has been addressed by Horváth et al. (2010) for AR(1)-processes and by Horváth and Reeder (2011) for more general processes by testing "classical" hypotheses (of the type H_0 versus H_1 described at the beginning of Section 3.1). In this Section we discuss how one can adapt the previous techniques to the detection of a relevant change in the operator S. The related, but easier case of comparing two operators, say $S^{(1)}$ and $S^{(2)}$ from independent samples is briefly discussed in Remark 5.3 below.

To be precise consider the following regression model

$$Y_n \otimes X_n = S_n X_n \otimes X_n + \varepsilon_n \otimes X_n \qquad n = 1, ..., N, \tag{5.1}$$

where $S^{(1)} := S_1 = S_2 = ... = S_{\lfloor \theta N \rfloor}$, $S^{(2)} := S_{\lfloor \theta N \rfloor + 1} = ... = S_N$ and $\theta \in (0, 1)$ determines the location of the change point and is unknown. We assume that $\{(X_n, \varepsilon_n)\}_{n \in \mathbb{Z}}$ is a stationary time series of regressors and errors, which satisfies the Assumptions 3.1(1)-(5) in Section 3.2. The two hypotheses of no relevant change at θ in the slope operator and of no relevant change in the predictive properties of S are defined by

$$H_0^\Delta : |||S^{(1)} - S^{(2)}|||^2 \leq \Delta \qquad H_1^\Delta : |||S^{(1)} - S^{(2)}|||^2 > \Delta, \tag{5.2}$$

$$^{\text{pred}}H_0^\Delta : \mathbb{E}\|S^{(1)}X - S^{(2)}X\|^2 \leq \Delta \qquad ^{\text{pred}}H_1^\Delta : \mathbb{E}\|S^{(1)}X - S^{(2)}X\|^2 > \Delta. \tag{5.3}$$

Before continuing we point out an important difference to change point analysis based on testing classical hypotheses (that is $\Delta = 0$): Suppose a change in the slope operator is detected by a traditional change point test, but would be considered irrelevant in the sense of the hypotheses (5.2) for some small $\Delta > 0$. In

this situation it might be reasonable to reconstruct the slope S using all of the data, instead of considering two estimates from the data before and after the estimated change point. On the one hand this would introduce a (small) bias in the estimation, but on the other hand this increase could be compensated by a significant reduction of the variance.
In the following discussion let $\hat{\theta}$ denote an estimator of the change point (see Remark 5.3(1) below for a concrete example). We define the sequential estimators for the covariance operator

$$\hat{\Gamma}_N^{(1)}[\xi] = \frac{1}{N\hat{\theta}} \sum_{n=1}^{\lfloor \xi\hat{\theta}N \rfloor} X_n \otimes X_n \qquad \text{and} \qquad \hat{\Gamma}_N^{(2)}[\xi] = \frac{1}{N(1-\hat{\theta})} \sum_{n=\hat{\theta}N+1}^{\lfloor \xi(1-\hat{\theta})N \rfloor} X_n \otimes X_n \ . \tag{5.4}$$

The eigenvalues (in non-increasing order) and their corresponding eigenfunctions are denoted by $\hat{\lambda}_1^{(j)}[\xi] \geq \hat{\lambda}_2^{(j)}[\xi] \geq \ldots$ and $\hat{e}_1^{(j)}[\xi], \hat{e}_2^{(j)}[\xi], \ldots$, respectively ($j = 1, 2$). As before, we consider for $k \in \mathbb{N}$ the regularized inverse of the operator $\hat{\Gamma}_N^{(j)}[\xi]$, as well as the projection on the first k empirical eigenfunctions as

$$\hat{\Gamma}_k^{\dagger,(j)}[\xi] = \sum_{i=1}^{k} \frac{1}{\hat{\lambda}_i^{(j)}[\xi]} \hat{e}_i^{(j)}[\xi] \otimes \hat{e}_i^{(j)}[\xi] \qquad \text{and} \qquad \hat{\Pi}_k^{(j)}[\xi] = \sum_{i=1}^{k} \hat{e}_i^{(j)}[\xi] \otimes \hat{e}_i^{(j)}[\xi].$$

By virtue of the regularized inverse operators, we can now define the slope estimates, as

$$\hat{S}_N^{(1)}[\xi] = \frac{1}{N\hat{\theta}} \sum_{n=1}^{\lfloor \xi\hat{\theta}N \rfloor} Y_n \otimes X_n \hat{\Gamma}_k^{\dagger,(1)}[\xi] \qquad \text{and} \qquad \hat{S}_N^{(2)}[\xi] = \frac{1}{N(1-\hat{\theta})} \sum_{n=N\hat{\theta}+1}^{\lfloor \xi(1-\hat{\theta})N \rfloor} Y_n \otimes X_n \hat{\Gamma}_k^{\dagger,(2)}[\xi] \tag{5.5}$$

and propose to reject the null hypothesis in (5.2) whenever

$$\hat{W}_N^{\mathrm{cp}}(\Delta) := \frac{\sqrt{N}\left(|\!|\!| \hat{S}_N^{(1)} - \hat{S}_N^{(2)} |\!|\!|^2 - \Delta \right)}{\hat{V}_N^{\mathrm{cp}}} > q_{1-\alpha}. \tag{5.6}$$

where the denominator $\hat{V}_N^{\mathrm{cp}}$ is defined as

$$\hat{V}_N^{\mathrm{cp}} := \left\{ \int_I \xi^4 \left(|\!|\!| \hat{S}_N^{(1)}[\xi] - \hat{S}_N^{(2)}[\xi] |\!|\!|^2 - |\!|\!| \hat{S}_N^{(1)} - \hat{S}_N^{(2)} |\!|\!|^2 \right)^2 d\nu(\xi) \right\}^{1/2}, \tag{5.7}$$

and $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the distribution of the random variable W defined in (3.14). In order to test for relevant predictive differences we define $\hat{W}_N^{\mathrm{cp,pred}}(\Delta)$ in the same way as $\hat{W}_N^{\mathrm{cp}}(\Delta)$, where we replace all instances of $S^{(j)}[\xi]$ by $S^{(j)}[\xi]\hat{\Gamma}_N^{(j)}[\xi]^{1/2}$. This gives us the test decision for a relevant change in prediction

$$\hat{W}_N^{\mathrm{cp,pred}}(\Delta) > q_{1-\alpha}. \tag{5.8}$$

For the statement of the main results of this section we require the consistency of the change point estimator $\hat{\theta}$, such that the amount of missclassified data is asymptotically negligible.

Assumption 5.1. (Consistency of $\hat{\theta}$): $\qquad \hat{\theta} = \theta + o_{\mathbb{P}}(1/\sqrt{N})$.

The following result shows that the decision rules (5.6) and (5.8) define consistent tests for the hypotheses (5.2) and (5.3), respectively and have asymptotic level α. In its formulation we understand that a postulated assumption applies to each operator before and after the change point.

Theorem 5.2. *Suppose that the Assumptions 3.1(1)-(6) and Assumption 5.1 hold.*

a) If additionally Assumption 3.1(7) holds, then the long-run variance $(\tau^{\mathrm{cp}})^2$ *of*

$$\sqrt{N}\left(|||\hat{S}_N^{(1)} - \hat{S}_N^{(2)}|||^2 - |||S^{(1)} - S^{(2)}|||^2\right) \xrightarrow{d} \mathcal{N}(0, (\tau^{\mathrm{cp}})^2)$$

exists. If τ^{cp} *is positive, then the decision rule in* (5.7) *yields a consistent, asymptotic level-*α *test for the hypothesis* (5.2) *of no relevant change in the slope.*

b) If additionally Assumption 4.1 holds, then the long-run variance $(\tau^{\mathrm{cp,pred}})^2$ *of*

$$\sqrt{N}\left(|||\hat{S}_N^{(1)}(\hat{\Gamma}_N^{(1)})^{1/2} - \hat{S}_N^{(2)}(\hat{\Gamma}_N^{(2)})^{1/2}|||^2 - |||(S^{(1)} - S^{(2)})\Gamma^{1/2}|||^2\right) \xrightarrow{d} \mathcal{N}(0, (\tau^{\mathrm{cp,pred}})^2)$$

exists. If $\tau^{\mathrm{cp,pred}}$ *is positive, then the decision rule in* (5.8) *yields a consistent, asymptotic level-*α *test for the hypothesis* (5.3) *of no relevant change in the prediction.*

It is possible to give explicit formulas for τ^{cp} and $\tau^{\mathrm{cp,pred}}$, which are similar to those in Proposition 3.3 and Theorem 4.2 above, but we omit them to avoid redundancy. We conclude this section with a brief remark concerning the change point estimator $\hat{\theta}$ and two sample testing.

Remark 5.3.

(1) There are many ways of defining an estimator for the location θ of the change point. As an example we consider a standard change point estimator from the observations $Y_1 \otimes X_1 \ldots, Y_n \otimes X_n$ based on the CUSUM-principle (note that any change in the slope operator S_n in model (5.1) manifests itself in the product $S_n\Gamma$). To be precise we define

$$\hat{\theta} := \frac{1}{N} \operatorname{argmax}\{f(M) : 1 < M < N\}, \tag{5.9}$$

where the function f is given by

$$f(M) := \frac{M}{N}\left(1 - \frac{M}{N}\right)\left\|\frac{1}{M}\sum_{n=1}^{M} Y_n \otimes X_n - \frac{1}{N-M}\sum_{n=M+1}^{N} Y_n \otimes X_n\right\|^2, \quad M = 2, \ldots, N-1$$

It then follows from Corollary 1 in Hariz et al. (2007) that

$$|\hat{\theta} - \theta| = \mathcal{O}_{\mathbb{P}}(1/N)$$

if the condition

$$\sum_{q,r} \sqrt{\mathbb{E}\langle S^{(j)}X \otimes X, f_r \otimes e_q\rangle^2} < \infty$$

holds for some orthonormal basis $\{f_r\}_{r\geq 1}$ of H_2. In this case Assumption 5.1 is satisfied for the estimator in (5.9).

(2) It is easy to see that the test formulated in this section can be applied (without the change point estimation) to the case of two independent samples of size N_1 and N_2. In this case we set $N = N_1 + N_2$ and replace $\hat{\theta}N$ and $(1-\hat{\theta})N$ in the definitions (5.4) and (5.5) by N_1 and N_2. The details are omitted for the sake of brevity. Tests for relevant differences between the slopes of two functional linear regression models may be of interest e.g. in cases where the behavior of contemporary individuals at different geographical locations is compared.

6 Finite sample properties

In this section we investigate the finite sample properties of the spectral reconstructions and the self-normalized tests by means of a simulation study. We restrict ourselves to the one-sample cases presented in Sections 3 and 4, even though non-reported simulations suggest similar performance for two sample cases and change point scenarios. Following Benatia et al. (2017) we consider $(H_1, \mu_1) = (H_2, \mu_2) = (L^2([0,1], \mu)$, where μ is the uniform distribution on the points $\{0, 1/50, 2/50..., 1\}$, which may be regarded as a discretized version of the Lebesgue measure. We consider two scenarios of dependence: i.i.d. observations and dependent observations, where regressors and errors are generated by $AR(1)$ processes.

6.1 Inference for the location of S

Recall the regression model in (1.1). In the case of i.i.d. observations, we generate the regressors as

$$X_n(t) = \frac{\Gamma(A_n + B_n)}{\Gamma(A_n)\Gamma(B_n)} t^{A_n}(1-t)^{B_n} + Z_n, \tag{6.1}$$

(shifted β-densities) where A_n, B_n are independent, uniformly distributed on the interval $[2, 5]$ and Z_n is an independent, standard normal shift. Notice that the regressor functions $X_1, X_2, \ldots$ are not centred, and hence we include an empirical centering in all statistics (see Remark 3.8). The error functions ε_n are i.i.d. realizations of an Ornstein–Uhlenbeck process, with zero mean, variance parameter and mean reversion rate equal to one. Note that ε_n is a stationary, centered, Gaussian process, which is the solution of the stochastic differential equation $d\varepsilon(t) = -\varepsilon(t)dt + \sigma d\mathbb{B}(t)$, where $\mathbb{B}$ is a standard Brownian motion and $\sigma = 1$. Some typical paths of these regressors and errors are depicted in Figure 1.

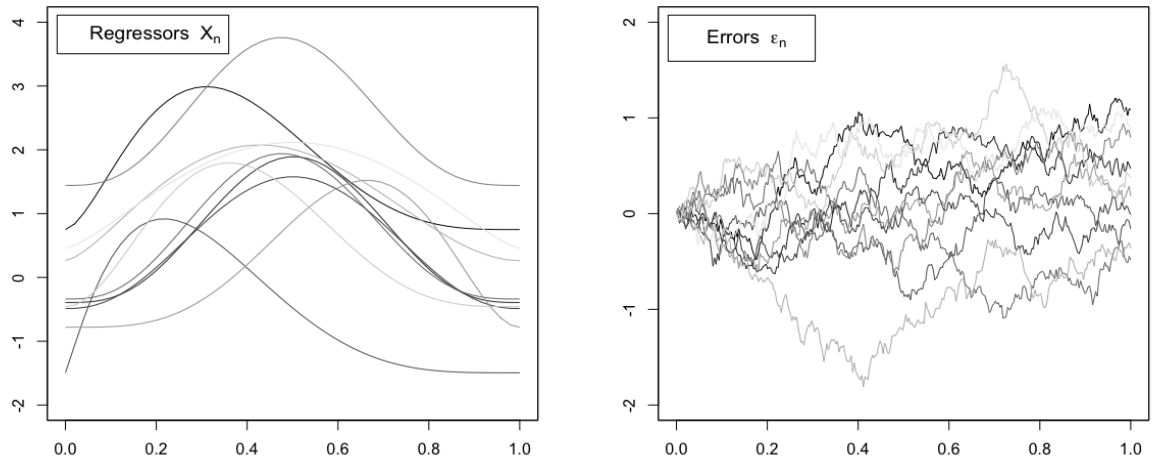


Figure 1: *Realizations of regressors (left) and errors (right), both in the i.i.d. case.*

In the case of dependent observations we generate both regressors and errors by $AR(1)$ processes, with parameter $\rho = 0.6$, that is

$$\tilde{X}_n = \rho \tilde{X}_{n-1} + X_n, \qquad \tilde{\varepsilon}_n = \rho \tilde{\varepsilon}_{n-1} + \varepsilon_n,$$

where the random variables X_n and ε_n are i.i.d. random functions, generated as in the independent case (see equation (6.1) and following discussion). In all simulations a burn in period of 200 repetitions is used. Finally we turn to the operators S_0 and S, both of which are integral operators, defined as

$$Sf \mapsto \int_0^1 \varphi_S(s,\cdot) f(s)\mu(s), \qquad \text{and} \qquad S_0 f \mapsto \int_0^1 \varphi_{S_0}(s,\cdot) f(s) d\mu(s),$$

pointwise for a function $f \in (L^2[0,1],\mu)$, where φ_S and φ_{S_0} are the integral kernels corresponding to S and S_0. We first define the benchmark kernel φ_{S_0} as in Benatia et al. (2017), by

$$\varphi_{S_0}(s,t) = 1 - |s-t|^2$$

and then the slightly more complex regression kernel φ_S as

$$\varphi_S(s,t) = 1 - 4/5|s-t|^2 + 1/5\cos(|s-t|/5).$$

In Figure 2 we plot the two kernel functions, to illustrate their shape differences. The difference between the kernels can be asessed by the relative deviation measure

$$1 - \frac{|||S - S_0|||^2}{|||S_0|||^2} = 1 - \frac{\int_0^1 \left[\varphi_S(s,t) - \varphi_{S_0}(s,t)\right]^2 ds\, dt}{\int_0^1 \varphi_{S_0}(s,t)^2 ds\, dt} \approx 0.032$$

(since $|||S|||^2 \approx |||S_0|||^2$ it does not matter by which norm we normalize).

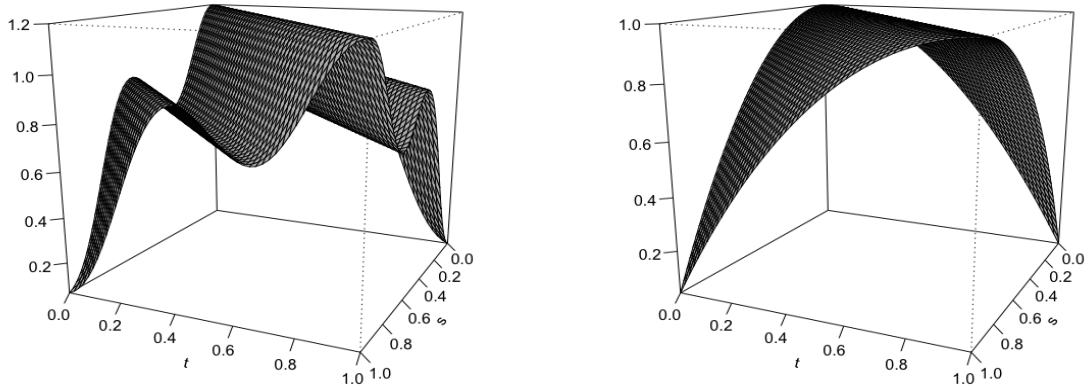


Figure 2: *Image of the two integration kernels, plotted as surfaces. Left: φ_S. Right: φ_{S_0}.*

Before we consider the estimation problem, it is reasonable to investigate the complexity of the two slopes S and S_0, relative to the principal components of the operator Γ. For this purpose we consider the measure of relative explanation

$$\frac{|||S\Pi_k - S_0\Pi_k|||^2}{|||S - S_0|||^2}, \tag{6.2}$$

which varies in the interval $[0, 1]$ and is increasing in k. A value of 1 means that $S - S_0$ acts exclusively on $\text{span}\{e_1, \ldots, e_k\}$, whereas a value of 0 implies that $S - S_0$ lives on the orthogonal complement. A rapid increase in k corresponds to a high degree of smoothness in the sense of Assumption 3.1(1) and hence to a small bias. However, smoothness of the slopes is not enough, since one also has to be able to estimate the principal components of Γ properly. This corresponds to eigenvalues $\lambda_1, \lambda_2, \ldots$ of Γ (and eigengaps), which are not too small.

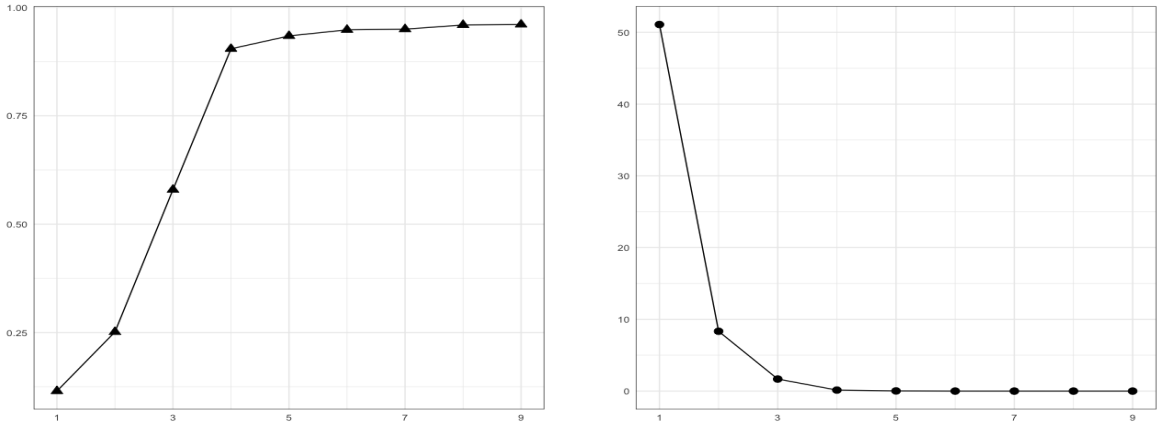
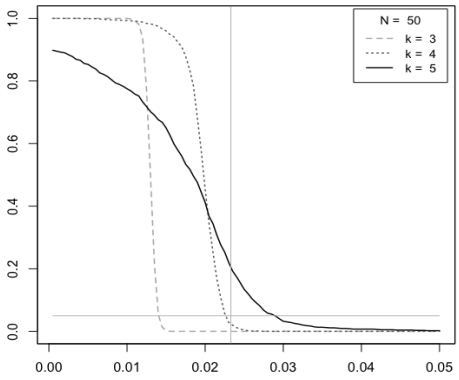


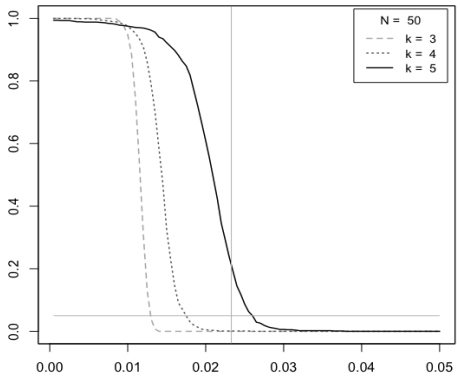
Figure 3: *Left: The measure* (6.2) *as a function of* k. *Right: First* 9 *eigenvalues of the operator* Γ.

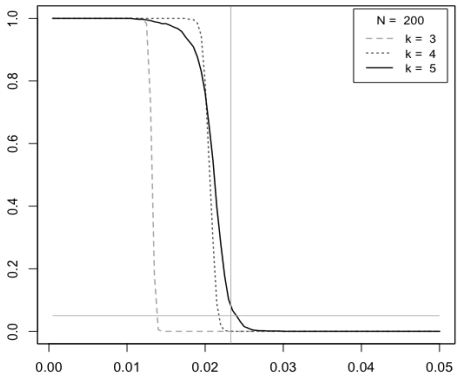
In left part of Figure 3 we display the measure (6.2) as a function of k and observe a quick increase for $k \leq 5$ up to 95%. In the right part of the figure we present the decaying eigenvalues of the operator Γ (in the case of independent variables). Here we observe a fast decay followed by a sharp elbow. The first three eigengaps are rather large, but afterwards the distinction becomes increasingly difficult. Indeed even for a large sample size of $N = 1000$, the recovery of more than 5 eigenfunctions is somewhat unstable, resulting in inflated rejection probabilities at the boundary of the hypothesis. Together these considerations suggest that choices of k between 4 and 5 are optimal, depending on the sample size N.

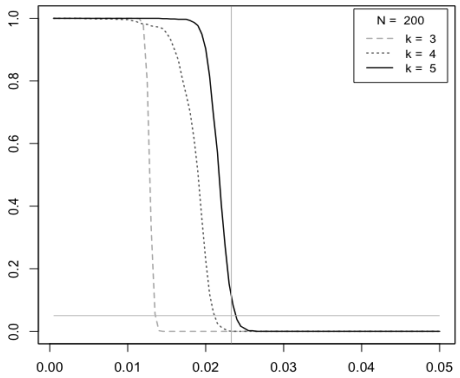
Throughout this section all empirical results are based on 1000 simulation runs. In order to investigate the power of the test (3.15) for the relevant hypotheses (3.2) we consider the sample sizes of $N = 50, 200, 500$ and $k = 3, 4, 5$ (note that $k = 3$ is rather small for practical inference and only included to illustrate aspects of the bias-variance trade-off). The measure ν in the definition of the normalizer (3.13) is the uniform distribution on the set $\{1/5, 2/5, 3/5, 4/5\}$. Simulations for other measures, which are not reported for the sake of brevity, suggest that the number of points does not have a large or systematic impact on the results. In Figure 4 we display the rejection probability of the self-normalized test (3.15) as a function of the threshold Δ in the hypothesis (3.2). A vertical grey line indicates the true value of $|||S - S_0|||^2 \approx 0.023$

and corresponds to the boundary of the hypotheses, while the grey horizontal line shows the nominal level α, which is chosen as $\alpha = 0.05$. The left column shows the results for the i.i.d. case, while the results for the dependent case can be found in the right column. The plots can be evaluated as follows:









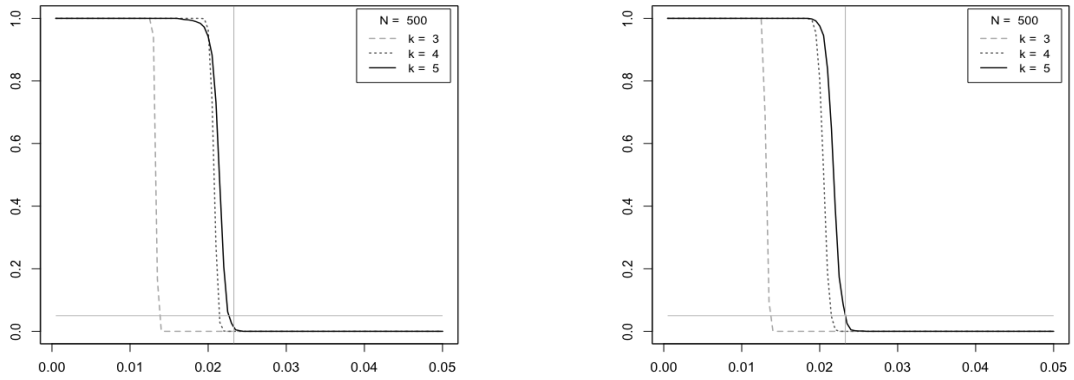


Figure 4: *Rejection probabilities (y-axis) of the test* (3.15) *for different choices of* Δ *(x-axis). Different sizes of* N *and* k *are displayed, for i.i.d. observations (left) and dependent observations (right). The level* $\alpha = 0.05$ *is indicated by the horizontal grey line, and the true size of* $|||S - S_0|||^2$ *by the vertical line.*

If $\Delta < |||S - S_0|||^2 \approx 0.023$ (left of the vertical line) we operate under the alternative (see (3.2)) and expect high rejection probabilities. At the boundary of the hypotheses corresponding to the vertical line we expect to approximate α, whereas for larger values of Δ the asymptotic theory developed in Section 3 suggests that the rejection probability tends to 0; see Remark 3.8(2). Because the test statistic is a monotone function of Δ, the rejection probability decreases monotonically in Δ; see Remark 3.8(3).

Before we evaluate the specific performance of the test (3.15), we illustrate in Figure 4 some general features of the linear inverse problem. Evidently the rejection curves exhibit the bias variance trade-off, which occurs at the level of k. For $k = 3$ we observe a substantial bias, which we would expect, in view of Figure 3 (left). It diminishes rapidly for $k = 4$ and $k = 5$. In view of our discussion of (3.9) we should understand the left shifts of the rejection curves as a result of the bias, which makes the test somewhat conservative. The upside of smaller values of k is the accompanying small variance, manifest in the rapid decay of the rejection curves. For larger values of k the variance increases and this may result in inflated rejection probabilities at the boundary (e.g. for $N = 50$ and the too large choice $k = 5$).

With regard to the statistical properties of the test presented in (3.15), we observe a reasonable level approximation at the boundary, even for sample sizes as small as 50 in the independent scenario. Dependence in the observations leads to worse performance, particularly for samples as small as $N = 50$, whereas the effect for $N = 200, 500$ is minute. The power of the test is for independent observations even for $N = 50$ acceptable and for larger samples, we observe rapid improvements, even for greater values of k. In the dependent case for $N = 200, 500$ we observe a good level approximation at the boundary and high rejection probabilities under the alternative. Interestingly the bias-variance trade-off sometimes implies that while some k leads to the optimal level approximation at the boundary and thus high power close to the hypothesis, for larger distances smaller values of k perform better, because the variance is smaller. This effect is reflected by crossing rejection curves. Finally, we notice that in view of the sometimes abrupt

change in variance and bias for two successive values of k, in practice a soft threshold might be considered, for a more nuanced trade-off.

6.2 Inference for relevant prediction errors

We consider the set-up described in the previous section to investigate deviation in the predictive performance of S and S_0. We begin by considering the smoothed kernels $S\Gamma^{1/2}$ and $S_0\Gamma^{1/2}$, which are depicted in Figure 5 (for Γ corresponding to the i.i.d. case). Even though they bear some resemblance to their originals (see Figure 2), we observe a high smoothing effect caused by the application of $\Gamma^{1/2}$.

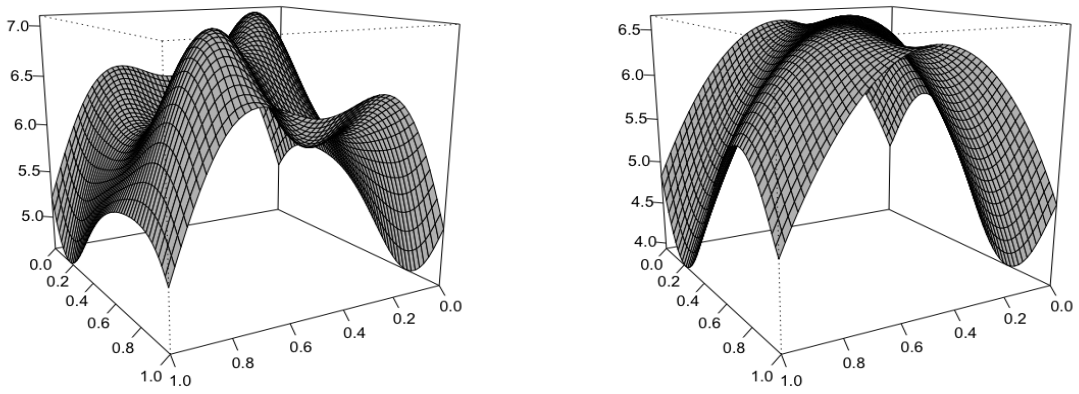


Figure 5: *Image of the two integration kernels, plotted as surfaces. Left:* $\varphi_{S\Gamma^{1/2}}$. *Right:* $\varphi_{S_0\Gamma^{1/2}}$.

As a consequence of the smoothing effect, we expect that the relative explanation should be higher than for the non-smoothed operators. This is in fact what we see in Figure 6 (left), where we have plotted the relative explanation measure

$$\frac{|||S\Gamma^{1/2}\Pi_k - S_0\Gamma^{1/2}\Pi_k|||^2}{|||S\Gamma^{1/2} - S_0\Gamma^{1/2}|||^2}. \tag{6.3}$$

We see that the first principal component already covers more than 75% of the norm, for $k = 3$ the relative explanation is about 99% and for $k = 4$ it has reached almost 100%. Compared to the explanation for the non-smoothed kernel in Figure 3 (left) this is a very rapid increase and it suggests the use of smaller values for k. Notice that this matches our theoretical results in Sections 3 and 4 (compare Assumptions (3.1)(7) and 4.1)), which suggest higher k for the recovery of the slope and smaller k for the purpose of prediction. On the right side of Figure 6 we display the smoothing kernel corresponding to the operator $\Gamma^{1/2}$ in the case of i.i.d. observations.

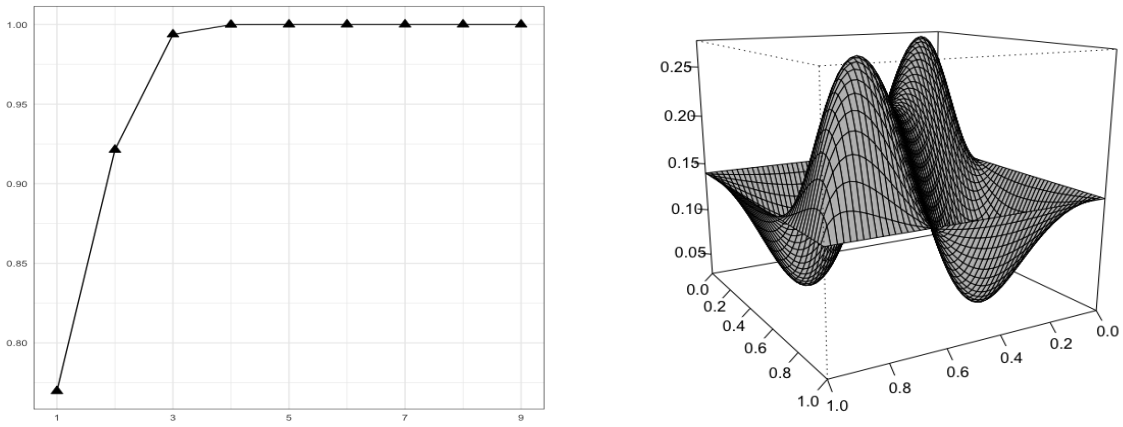
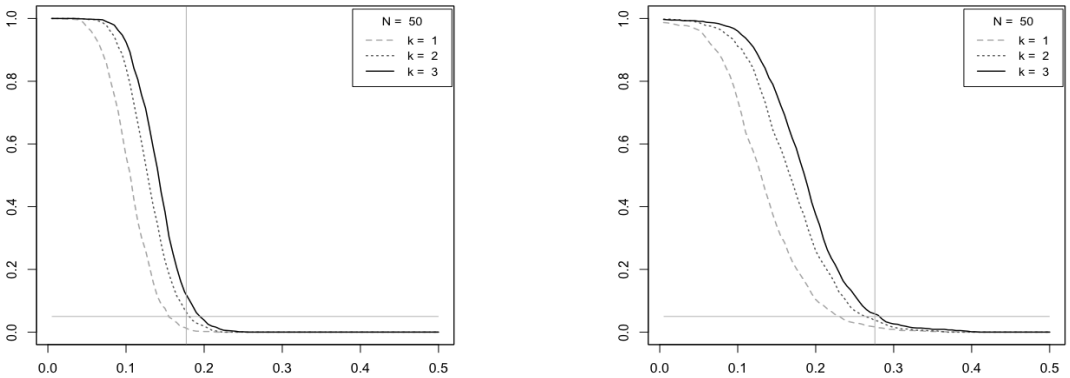


Figure 6: *Left: Relative explanation measure defined in* (6.3) *as a function of* k. *Right: Kernel of the operator* $\Gamma^{1/2}$ *(for i.i.d. observations).*

We now proceed to the application of the statistical test (4.5), presented in Section 4 for the hypothesis (4.2). As in Section 6.1 we consider sample sizes $N = 50, 200, 500$ and parameter choices $k = 1, 2, 3$, both for i.i.d. samples (left part of the figures) and dependent samples (right part of the figures). For details on the model as well as the dependence we refer to Section 6.1. The measure ν in the normalizer $\hat{V}_N^{\mathrm{pred}}$ (see (4.4)) is again chosen to be the uniform measure on $\{1/5, 2/5, 3/5, 4/5\}$ and the level of the test is $\alpha = 0.05$. All simulations are based on 1000 repetitions. Notice that the absolute deviation (vertical grey line) differs in the case of independent and dependent data, since the operator Γ is different in the dependent and independent case.



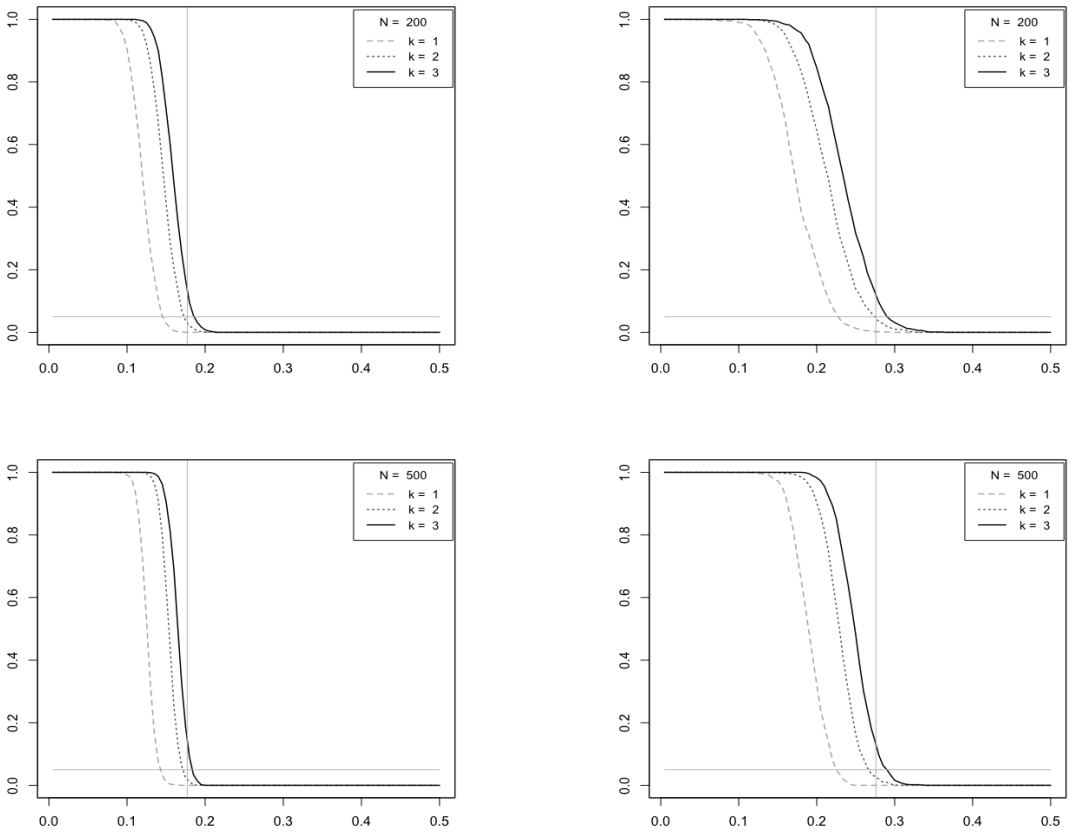


Figure 7: *Rejection probabilities (y-axis) of the test* (4.5) *for different choices of* Δ *(x-axis). Different sizes of* N *and* k *are displayed, for i.i.d. observations (left) and dependent observations (right). The level* $\alpha = 0.05$ *is indicated by the horizontal grey line, and the true size of* $|||S\Gamma^{1/2} - S_0\Gamma^{1/2}|||^2$ *by a vertical line.*

The numerical results confirm the theoretical findings in Section 4. We observe a good approximation of the level at the boundary of the hypothesis, both for dependent and independent data and accordingly high rejection probabilities under the alternative. The smoothing parameter is chosen smaller than in the case of prediction, which corresponds to the smaller bias in the case of prediction. In contrast to Section 6.1 we do not see pronounced crossing of the power curves for different k, such that better level approximation automatically translates into higher overall power. This also is an effect of the relatively small bias in the case of prediction.

References

Andersson, J. and J. Lillestøl (2010). Modeling and forecasting electricity consumption by functional data analysis. *Journal of Energy Markets 3*(1), 3–15.

Aston, J. A. and C. Kirch (2012a). Detecting and estimating changes in dependent functional data. *J. Multivariate Anal. 109*, 204–220.

Aston, J. A. and C. Kirch (2012b). Evaluating stationarity via change-point alternatives with applications to fMRI data. *Ann. Appl. Stat. 6*(4), 1906–1948.

Aue, A., G. Rice, and O. Sönmez (2020). Structural break analysis for spectrum and trace of covariance operators. *Environmetrics 31*(1), e2617. e2617 env.2617.

Babii, A. (2020). Honest confidence sets in nonparametric IV regression and other ill-posed models. *Econometric Theory 36*(4), 658–706.

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

Benko, M., W. Härdle, and A. Kneip (2009). Common functional principal components. *Ann. Statist. 37*(1), 1–34.

Berkes, I., R. Gabrys, L. Horváth, and P. Kokoszka (2009). Detecting changes in the mean of functional observations. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 71*(5), 927–946.

Berkes, I., L. Horváth, and G. Rice (2013). Weak invariance principles for sums of dependent random functions. *Stochastic Process. Appl. 123*(2), 385–403.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *J. Amer. Statist. Assoc. 33*(203), 526–536.

Bissantz, N., H. Dette, and K. Proksch (2012). Model checks in inverse regression models with convolution-type operators. *Scand. J. Stat. 39*(2), 305–322.

Bissantz, N. and H. Holzmann (2008). Statistical inference for inverse problems. *Inverse Problems 24*(3), 034009.

Bonner, S., N. Newlands, and N. Heckman (2014). Modeling regional impacts of climate teleconnections using functional data analysis. *Environmental and Ecological Statistics 21*, 1–26.

Bücher, A. and I. Kojadinovic (2013). A dependent multiplier bootstrap for the sequential empirical copula process under strong mixing. *Bernoulli 22*(2), 927–968.

Cardot, H., F. Ferraty, A. Mas, and P. Sarda (2003). Testing hypotheses in the functional linear model. *Scand. J. Stat. 30*(1), 241–255.

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Statist. Sinica 13*(3), 571–591.

Cardot, H., A. Goia, and P. Sarda (2004). Testing for no effect in functional linear regression models, some computational approaches. *Comm. Statist. Simulation Comput. 33*, 179–199.

Cardot, H., A. Mas, and P. Sarda (2007). CLT in functional linear regression models. *Probab. Theory Related Fields 138*, 325–361.

Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems 24*(3), 034004.

Constantinou, P., P. Kokoszka, and M. Reimherr (2017). Testing separability of space–time functional processes. *Biometrika 104*(2), 425–437.

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

Dehling, H. (1983). Limit theorems for sums of weakly dependent banach space valued random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 63*(3), 393—432.

Dehling, H., T. Mikosch, and M. Sørensen (2002). *Empirical process techniques for dependent data.* Birkhäuser.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Statist. 27*(3), 1012 – 1040.

Dette, H., G. Dierickx, and T. Kutta (2021). Quantifying deviations from separability in space-time functional processes. `https://arXiv:2003.12126`.

Dette, H., K. Kokot, and S. Volgushev (2020). Testing relevant hypotheses in functional time series via self-normalization. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 82*(3), 629–660.

Dette, H. and T. Kutta (2021). Detecting structural breaks in eigensystems of functional time series. *Electron. J. Stat. 15*(1), 944–983.

Dunford, N. and J. T. Schwartz (1958). *Linear operators. Part I: General theory.* New York, Interscience Publishers.

Engl, H. W., M. Hanke, and A. Neubauer (1996). *Regularization of inverse problems.* Dordrecht: Kluwer Academic Publishers Group.

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

Hariz, S., J. Wylie, and Q. Zhang (2007). Optimal rate of convergence for nonparametric change-point estimators for nonstationary sequences. *Ann. Statist. 35*, 1802–1826.

Herrndorf, N. (1983). The invariance principle for ϕ-mixing sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 63*(1), 97–108.

Hilgert, N., A. Mas, and N. Verzelen (2013). Minimax adaptive tests for the functional linear model. *Ann. Statist. 41*(2), 838 – 869.

Hörmann, S. and L. Kidzinski (2012). A note on estimation in Hilbertian linear models. *Scand. J. Stat. 42*(1), 43–62.

Hörmann, S. and P. Kokoszka (2010). Weakly dependent functional data. *Ann. Statist. 38*(3), 1845–1884.

Horváth, L., M. Hušková, and P. Kokoszka (2010). Testing the stability of the functional autoregressive process. *J. Multivariate Anal. 101*(2), 353–367.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. New York: Springer Series in Statistics.

Horváth, L., P. Kokoszka, and R. Reeder (2011). Estimation of the mean of functional time series and a two sample problem. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 75*, 103–122.

Horváth, L. and R. Reeder (2011). Detecting changes in functional linear models. *J. Multivariate Anal. 111*, 310–334.

Imaizumi, M. and K. Kato (2018). PCA-based estimation for functional linear regression with functional responses. *J. Multivariate Anal. 163*, 15–36.

Imaizumi, M. and K. Kato (2019). A simple method to construct confidence bands in functional linear regression. *Statist. Sinica 29*(4), 2055–2081.

Jarušková, D. (2013). Testing for a change in covariance operator. *J. Statist. Plann. Inference 143*(9), 1500–1511.

Kokoszka, P. (2012). Dependent functional data. *Int Sch Res Notices Probability and Statistics 2012*.

Kokoszka, P. and M. Reimherr (2013). Asymptotic normality of the principal components of functional time series. *Stochastic Process. Appl. 123*(5), 1546–1562.

Kong, D., A.-M. Staicu, and A. Maity (2016). Classical testing in functional linear models. *J. Nonparametr. Stat. 28*(4), 813–838. PMID: 28955155.

Künsch, H. R. (1989). The jacknife and the bootstrap for general stationary observations. *Ann. Statist. 17*(3), 1217–1241.

Leung, R. C. W. and Y.-M. Tam (2021). A small-uniform statistic for the inference of functional linear regressions. `https://arxiv.org/abs/2102.10724`.

Liebl, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Stat. 7*(3), 1562–1592.

Merlevède, F., M. Peligrad, and S. Utev (2006). Recent advances in invariance principles for stationary sequences. *Probab. Surv. 3*, 1–36.

Moricz, F. A., R. J. Serfling, and W. F. Stout (1982). Moment and probability bounds with quasi-superadditive structure for the maximum partial sum. *Ann. Probab. 10*(4), 1032–1040.

Politis, D. and J. Romano (1994). The stationary bootstrap. *J. Amer. Statist. Assoc. 89*, 1303–1313.

Qiao, X., S. Guo, and G. M. James (2019). Functional graphical models. *J. Amer. Statist. Assoc. 114*, 211–222.

Ramsay, J. O. and B. W. Silverman (1997). *Functional Data Analysis.* Berlin, Springer.

Samur, J. D. (1984). Convergence of sums of mixing triangular arrays of random vectors with stationary rows. *Ann. Probab. 12*(2), 390–426.

Scheipl, F. and S. Greven (2016). Identifiability in penalized function-on-function regression models. *Electron. J. Stat. 10*(1), 495–526.

Shao, X. (2015). Self-normalization for time series: A review of recent developments. *J. Amer. Statist. Assoc. 110*, 1797–1817.

Shin, H. and S. Lee (2016). An rkhs approach to robust functional linear regression. *Statist. Sinica 26*, 255–272.

Sørensen, H., J. Goldsmith, and L. Sangalli (2013). An introduction with medical applications to functional data analysis. *Stat. Med. 32*, 5222–5240.

Stöhr, C., J. Aston, and C. Kirch (2021). Detecting changes in the covariance structure of functional time series with application to fmri data. *Econom. Stat.*, 44–62.

Utev, S. A. (1991). On the central limit theorem for φ-mixing arrays of random variables. *Theory Probab. Appl. 35*(1), 131–139.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes. With applications to statistics.* New York: Springer Series in Statistics.

Weidmann, J. (1980). *Linear Operators in Hilbert Spaces*, Volume 68 of *Graduate Texts in Mathematics*. Berlin, New York, Springer.

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist. 33*(6), 2873–2903.

Yuan, M. and T. Cai (2012). A reproducing kernel hilbert space approach to functional linear regression. *Ann. Statist. 38*(6), 3412–3444.

A Proofs and technical details

The Appendix is dedicated to the proofs of the theoretical results from Sections 3 to 5. We only show the weak convergence results from Section 3 explicitly, as those in Sections 4 and 5 are straightforward modifications. In the Appendix B we have collected some results concerning operators and their eigensystems. There (in Remark B.2) we also address the problem of the non-uniqueness of eigenfunctions, which is a technical issue, concerning the comparisons of eigenfunctions. Roughly speaking we always assume that the eigenfunctions e_i and their empirical counterparts (both unique up to sign) have "the same sign", in the sense that for all ξ the inequality $\|e_i - \hat{e}_i[\xi]\| \leq \|e_i + \hat{e}_i[\xi]\|$ holds. Notice that this technicality is of no concern in applications, as our test statistics always involve outer products of the form $\hat{e}_i[\xi] \otimes \hat{e}_i[\xi]$, which cancel the sign out.
Finally, we assume for notational simplicity that the sequential parameter ξ is contained in the interval $[1/2, 1]$ (any interval $[a, 1]$, $a > 0$, can be dealt with in the same way). In the remainder of this introduction we recall some useful identities for sequential operators and introduce a suitable sup-norm for them.

Let us revisit the sequential statistics defined in Section 3.3: Recall the definition of the sequential covariance estimate $\hat{\Gamma}_N[\xi]$ (see (3.10)), its eigenvalues and eigenfunctions $\hat{\lambda}_i[\xi], \hat{e}_i[\xi]$, $i \geq 1$, its regularized inverse $\hat{\Gamma}_k^\dagger[\xi]$ (see (3.11)), the projection $\hat{\Pi}_k[\xi]$, (see (3.11)) on $\text{span}\{\hat{e}_1[\xi], ..., \hat{e}_k[\xi]\}$, which can be expressed as $\hat{\Pi}_k[\xi] = \hat{\Gamma}_N[\xi]\hat{\Gamma}_k^\dagger[\xi]$ and finally the sequential estimate of S, denoted by $\hat{S}_N[\xi]$ (see (3.12)). Notice that for $\hat{S}_N[\xi]$ an analogue decomposition to (2.11) holds

$$\hat{S}_N[\xi] = S\hat{\Pi}_k[\xi] + U_N[\xi]\hat{\Gamma}_k^\dagger[\xi],$$

where the sequential residual term is defined as

$$U_N[\xi] := \frac{1}{N} \sum_{n=1}^{\lfloor N\xi \rfloor} \varepsilon_n \otimes X_n.$$

For fixed ξ, each of these statistics is defined as an element of a suitable Hilbert space ($\hat{\lambda}_i[\xi] \in \mathbb{R}$, $\hat{e}_i[\xi] \in H_1$, $\hat{\Gamma}_N[\xi], \hat{\Gamma}_k^\dagger[\xi], \hat{\Pi}_k[\xi] \in \mathcal{S}(H_1, H_1)$ and $\hat{S}_N[\xi] \in \mathcal{S}(H_1, H_2)$). Alternatively, we may also view each of these statistics as a bounded function in ξ, mapping from $[1/2, 1]$ into the respective Hilbert space. We make this notion more precise by defining the space of bounded functions:

Definition A.1. *Let $\mathcal{B}$ be a Banach space with norm $|||\cdot|||_{\mathcal{B}}$. Then we denote by*

$$\ell^\infty(\mathcal{B}) := \big\{ f : [1/2, 1] \to \mathcal{B} : \sup_{\xi \in [1/2,1]} |||f(\xi)|||_{\mathcal{B}} < \infty \big\},$$

the space of all bounded functions with range $\mathcal{B}$. Endowed with the sup-norm $\ell^\infty(\mathcal{B})$ is itself a Banach space.

In the sense of Definition A.1, we see that e.g. $\hat{\Gamma}_N[\cdot] \in \ell^\infty(\mathcal{S}(H_1, H_1))$. We conclude this part with the observation that the sequential covariance estimator $\hat{\Gamma}_N[\xi]$ is asymptotically close to the true one $\xi\Gamma$.

Theorem A.2. *(Dette et al. (2021)) Under the Assumptions (3.1)(2) and (3) it holds that*

$$\sup_{\xi \in [1/2,1]} |||\sqrt{N}(\hat{\Gamma}_N[\xi] - \xi\Gamma)||| = \mathcal{O}_{\mathbb{P}}(1).$$

We now give the proof of the main Theorem 3.5. To make the proof easier to comprehend, the discussion of various remainders is bundled in later lemmata.

A.1 Proof of Theorem 3.5

The proof consists of two steps: First we derive an asymptotic linearization of the test statistic. Subsequently we show weak convergence to a Brownian motion.

Using the identity $\|a\|^2 - \|b\|^2 = 2\langle a-b, b\rangle + \|a-b\|^2$ (which is a version of the third binomial formula for inner products), it follows that

$$\begin{aligned}&\xi\sqrt{N}\left\{|||\hat{S}_N[\xi] - S_0\hat{\Pi}_k[\xi]|||^2 - |||S\Pi_k - S_0\Pi_k|||^2\right\} \\ =&2\xi\sqrt{N}\langle \hat{S}_N[\xi] - S_0\hat{\Pi}_k[\xi] - S\Pi_k + S_0\Pi_k, S\Pi_k - S_0\Pi_k\rangle + \xi\sqrt{N}\mathcal{R}_1^2[\xi].\end{aligned} \tag{A.1}$$

Here the term $\mathcal{R}_1[\xi]$ is defined as

$$\begin{aligned}\mathcal{R}_1[\xi] :=&|||\hat{S}_N[\xi] - S_0\hat{\Pi}_k[\xi] - S\Pi_k + S_0\Pi_k||| \\ =&|||(S-S_0)(\hat{\Pi}_k[\xi] - \Pi_k) + U_N[\xi]\hat{\Gamma}_k^\dagger[\xi]|||.\end{aligned} \tag{A.2}$$

Using the triangle inequality we see that $\mathcal{R}_1[\xi] \leq |||(S-S_0)(\hat{\Pi}_k[\xi] - \Pi_k)||| + |||U_N[\xi]\hat{\Gamma}_k^\dagger[\xi]|||$. Recalling that $S-S_0$ can be rewritten as $(R-R_0)\Gamma^\beta$ for suitable Hilbert–Schmidt operators R, R_0 (see Assumption 3.1(1)), implies that the first term on the right is bounded by

$$|||(S-S_0)(\hat{\Pi}_k[\xi] - \Pi_k)||| \leq |||R-R_0|||\,|||\Gamma^\beta(\hat{\Pi}_k[\xi] - \Pi_k)|||_{\mathcal{L}}. \tag{A.3}$$

Now Lemma A.3, together with the bound (A.3) implies for an arbitrarily small $\epsilon > 0$, that

$$\mathcal{R}_1[\xi] = \mathcal{O}_{\mathbb{P}}(k^{(\gamma+1)/2+\epsilon}/\sqrt{N}). \tag{A.4}$$

Consequently $\xi\sqrt{N}\mathcal{R}_1^2[\xi] = o_{\mathbb{P}}(1)$ whenever ϵ is chosen small enough, i.e., $\epsilon \leq 2\delta$ (see Assumption 3.1(7)). We now focus on the non-vanishing term on the right of (A.1). It can be further decomposed into the sum $T[\xi] + \mathcal{R}_2[\xi]$ where

$$T[\xi] := 2\xi\sqrt{N}\left\langle (S-S_0)(\hat{\Pi}_k[\xi] - \Pi_k) + U_N[\xi]\hat{\Gamma}_k^\dagger[\xi],\ S - S_0\right\rangle \tag{A.5}$$

and $\mathcal{R}_2[\xi]$ is another remainder term, defined as

$$\mathcal{R}_2[\xi] := 2\xi\sqrt{N}\left\langle (S-S_0)(\hat{\Pi}_k[\xi] - \Pi_k) + U_N[\xi]\hat{\Gamma}_k^\dagger[\xi],\ (S-S_0)(\Pi_k - \mathrm{Id})\right\rangle.$$

By the Cauchy–Schwarz inequality one has

$$|\mathcal{R}_2[\xi]| \leq 2\xi\sqrt{N}\mathcal{R}_1[\xi]|||(S-S_0)(\Pi_k - \mathrm{Id})|||.$$

From (A.4) we know that $\sqrt{N}\mathcal{R}_1[\xi] = \mathcal{O}_{\mathbb{P}}(k^{\gamma/2+1/2+\epsilon})$ and according to our discussion of Assumption 3.1(7) (see Remark 3.2 (c)) $|||(S-S_0)(\Pi_k - \mathrm{Id})||| = \mathcal{O}(k^{-\gamma\beta})$. Hence it follows that

$$\mathcal{R}_2[\xi] = \mathcal{O}_{\mathbb{P}}\left(k^{\gamma/2+1/2+\epsilon-\gamma\beta}\right),$$

which is $o_{\mathbb{P}}(1)$ for a sufficiently small choice of ϵ, namely $\epsilon < (\gamma+1)/2$ where we used $\beta > 1 + 1/\gamma$ (again see discussion of Assumption 3.1(7) in Remark 3.2 (c)). We now analyze the non-vanishing term $T[\xi]$ defined in equation (A.5). Recall that by Assumption 3.1(1) two Hilbert–Schmidt operators R, R_0 exist, such that $S = R\Gamma^\beta$ and $S_0 = R_0\Gamma^\beta$. It thus follows that

$$\begin{aligned} T[\xi] =& 2\xi\sqrt{N}\left\langle (R-R_0)\Gamma^\beta(\hat{\Pi}_k[\xi] - \Pi_k) + U_N[\xi]\hat{\Gamma}_k^\dagger[\xi],\ (R-R_0)\Gamma^\beta \right\rangle \\ =& 2\xi\sqrt{N}\left\langle \left\{(R-R_0)\Gamma^\beta(\hat{\Pi}_k[\xi] - \Pi_k) + U_N[\xi]\hat{\Gamma}_k^\dagger[\xi]\right\}\Gamma^\beta,\ R-R_0 \right\rangle \end{aligned}$$

Notice that we have "shifted" the smoothing operator Γ^β from the right side of the inner product to the left. This shift can be validated by basic calculations. However it can be seen more easily as an application of the cyclical property for the trace of operators (see Horváth and Kokoszka (2012), Section 13.5 for details). Finally, we use Lemma A.5 to replace $\xi\Gamma^\beta(\hat{\Pi}_k - \Pi_k)\Gamma^\beta$ and $\xi U_N[\xi]\hat{\Gamma}_k^\dagger[\xi]\Gamma^\beta$ on the right, by their asymptotic linearizations, which yields $T[\xi] = T'[\xi] + o_{\mathbb{P}}(1)$, where

$$T'[\xi] = 2\sqrt{N}\left\langle (R-R_0)L(\hat{\Gamma}_N[\xi] - \xi\Gamma) + U_N[\xi]\Gamma^{\beta-1},\ R-R_0 \right\rangle. \tag{A.6}$$

Here L is a linear map defined in Definition A.4. Our proof up to this point implies the (asymptotic) stochastic linearization

$$\sqrt{N}\left\{ |||\hat{S}_N[\xi] - S_0\hat{\Pi}_k[\xi]|||^2 - |||S\Pi_k - S_0\Pi_k|||^2 \right\} = T'[\xi] + o_{\mathbb{P}}(1).$$

Now, having completed the linearization we still have to show that $T'[\xi]$ converges weakly to $\tau\mathbb{B}[\xi]$. By definition of $T'[\xi]$ (in (A.6)), it can be written as a real valued, sequential sum process of ϕ-mixing random variables. Therefore, we can apply the invariance principle from Corollary 2.6 in Herrndorf (1983) (where Condition (E) can be verified using our Assumption 3.1(2), with $\delta := \kappa/2$). This directly implies $T'[\xi] \xrightarrow{d} \tau\mathbb{B}[\xi]$, where τ^2 is the long-run variance of the statistic. It is now easy to calculate that τ^2 has the form postulated in (3.7), which concludes the proof. □

In the subsequent sections we show several bounds for the remainder terms, we have used throughout the proof of Theorem 3.5.

A.2 Bounds for $\mathcal{R}_1$

In the next Lemma A.3 we give orders of magnitude for $|||U_N[\xi]\hat{\Gamma}_k^\dagger[\xi]|||$ and $|||\Gamma^\beta(\hat{\Pi}_k[\xi] - \Pi_k)|||_{\mathcal{L}}$. Together with (A.3) (and the following part up to (A.4)) these imply that $\mathcal{R}_1$ satisfies the decay rate (A.4).

Lemma A.3. *Under the conditions of Theorem 3.5, it holds that*

i)
$$\sup_{\xi\in[1/2,1]} |||\Gamma^\beta(\hat{\Pi}_k[\xi] - \Pi_k)|||_{\mathcal{L}} = \mathcal{O}_{\mathbb{P}}(1/\sqrt{N}).$$

ii) For any $\epsilon > 0$
$$\sup_{\xi\in[1/2,1]} |||U_N[\xi]\hat{\Gamma}_k^\dagger[\xi]||| = \mathcal{O}_{\mathbb{P}}(k^{\gamma/2+1/2+\epsilon}/\sqrt{N}).$$

Proof.

i) We use (3.11) and the decomposition

$$\Gamma^\beta(\hat{\Pi}_k[\xi] - \Pi_k) = A_1[\xi] + A_2[\xi] + A_3[\xi],$$

where

$$A_1[\xi] := \sum_{i=1}^{k} \Gamma^\beta(\hat{e}_i[\xi] - e_i) \otimes e_i \tag{A.7}$$

$$A_2[\xi] := \sum_{i=1}^{k} \Gamma^\beta e_i \otimes (\hat{e}_i[\xi] - e_i) \tag{A.8}$$

$$A_3[\xi] := \sum_{i=1}^{k} \Gamma^\beta(\hat{e}_i[\xi] - e_i) \otimes (\hat{e}_i[\xi] - e_i). \tag{A.9}$$

We now show the desired rate for each term separately.

A_1) Recall the spectral decomposition of the covariance operator $\Gamma = \sum_{q\geq 1} \lambda_q e_q \otimes e_q$, which yields

$$A_1[\xi] = \sum_{i=1}^{k} \sum_{q\geq 1} \lambda_q^\beta \langle \hat{e}_i[\xi] - e_i, e_q\rangle e_q \otimes e_i.$$

Separating the terms where $i = q$ from the ones where $i \neq q$, we decompose

$$A_1[\xi] = A_{1,1}[\xi] + A_{1,2}[\xi],$$

with

$$A_{1,1}[\xi] := \sum_{i=1}^{k} \lambda_i^\beta \langle \hat{e}_i[\xi] - e_i, e_i\rangle (e_i \otimes e_i) = \sum_{i=1}^{k} \frac{-\lambda_i^\beta}{2} \left\|\hat{e}_i[\xi] - e_i\right\|^2 (e_i \otimes e_i), \tag{A.10}$$

$$A_{1,2}[\xi] := \sum_{i=1}^{k} \sum_{q\neq i} \lambda_q^\beta \langle \hat{e}_i[\xi] - e_i, e_q\rangle (e_q \otimes e_i). \tag{A.11}$$

Notice that we used identity $\langle \hat{e}_i[\xi] - e_i, e_i\rangle = -\left\|\hat{e}_i[\xi] - e_i\right\|^2/2$ in (A.10) (see Appendix B, Lemma B.3). We bound the operator norm of (A.10) and (A.11) individually.

$\boxed{A_{1,1}:}$ Recall that the operator norm of a diagonal operator equals its largest, absolute diagonal entry, i.e.

$$\left|\!\left|\!\left| A_{1,1}[\xi] \right|\!\right|\!\right|_{\mathcal{L}} = \max_{1\leq i\leq k} \frac{\lambda_i^\beta}{2} \left\|\hat{e}_i[\xi] - e_i\right\|^2. \tag{A.12}$$

Further, using the inequality

$$\|\hat{e}_i[\xi] - e_i\| \leq \frac{2\sqrt{2}\left|\!\left|\!\left| \hat{\Gamma}_N[\xi] - \xi\Gamma \right|\!\right|\!\right|_{\mathcal{L}}^2}{\xi \min(\lambda_{i-1} - \lambda_i, \lambda_i - \lambda_{i+1})},$$

(see Appendix B Lemma B.3) we have

$$\max_{1\leq i\leq k} \frac{\lambda_i^\beta}{2} \left\|\hat{e}_i[\xi] - e_i\right\|^2 \leq \max_{1\leq i\leq k} \frac{2\sqrt{2}\lambda_i^\beta \left|\!\left|\!\left| \hat{\Gamma}_N[\xi] - \xi\Gamma \right|\!\right|\!\right|_{\mathcal{L}}}{\min(\lambda_{i-1} - \lambda_i, \lambda_i - \lambda_{i+1})},$$

where we have used that $1/\xi \leq 2$. We now simplify the right side by Lemma A.7 part *iii*) and the fact that $\|\!|\hat{\Gamma}_N[\xi] - \xi\Gamma|\!\|_{\mathcal{L}} = \mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$ (Theorem A.2). Together these show that

$$\sup_{\xi\in[1/2,1]} \|\!|A_{1,1}[\xi]|\!\|_{\mathcal{L}} = \mathcal{O}_{\mathbb{P}}(1/N).$$

$\boxed{A_{1,2}:}$ Since the norm inequality $\|\!|\cdot|\!\|_{\mathcal{L}} \leq \|\!|\cdot|\!\|$ holds, it is enough to show that

$$\sup_{\xi\in[1/2,1]} \|\!|A_{1,2}[\xi]|\!\| = \mathcal{O}_{\mathbb{P}}(1/\sqrt{N}).$$

The identity (B.1) implies that, $\langle \hat{e}_i[\xi] - e_i, e_q\rangle = \langle \hat{\Gamma}_N[\xi] - \xi\Gamma, \hat{e}_i[\xi]\otimes e_q\rangle/(\hat{\lambda}_i[\xi] - \xi\lambda_q)$. We can now upper bound the Hilbert–Schidt norm of $A_{1,2}[\xi]$ as follows:

$$\begin{aligned}
\|\!|A_{1,2}[\xi]|\!\| &= \left(\sum_{i=1}^{k}\sum_{q\neq i} \lambda_q^{2\beta} \frac{\langle \hat{\Gamma}_N[\xi] - \xi\Gamma, \hat{e}_i[\xi]\otimes e_q\rangle^2}{(\hat{\lambda}_i[\xi] - \xi\lambda_q)^2}\right)^{1/2} \qquad (A.13)\\
&\leq \max_{i\leq k, q\neq i} \frac{\lambda_q^{\beta}}{|\hat{\lambda}_i[\xi] - \xi\lambda_q|} \left(\sum_{i=1}^{k}\sum_{q\neq i} \langle \hat{\Gamma}_N[\xi] - \xi\Gamma, \hat{e}_i[\xi]\otimes e_q\rangle^2\right)^{1/2}\\
&\leq \max_{i\leq k, q\neq i} \frac{\lambda_q^{\beta}}{|\hat{\lambda}_i[\xi] - \xi\lambda_q|} \|\!|\hat{\Gamma}_N[\xi] - \xi\Gamma|\!\|.
\end{aligned}$$

Here we have used that $\{\hat{e}_i[\xi]\}_{i\in\mathbb{N}}, \{e_q\}_{q\in\mathbb{N}}$ are ONBs and thus their products $\{\hat{e}_i[\xi]\otimes e_q\}_{i,q\in\mathbb{N}}$ form an ONB of the Hilbert–Schmidt operators (see Section 2.1). The fraction of the eigenvalues is uniformly of order $\mathcal{O}_{\mathbb{P}}(1)$ by Lemma A.7, part *ii*) whereas $\|\!|\hat{\Gamma}_N[\xi] - \xi\Gamma|\!\| = \mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$ by Theorem A.2.

Combining both estimates gives the (uniform) order $\mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$ for the term A_1.

A_2) As $\Gamma^{\beta} e_i = \lambda_i^{\beta} e_i$ we can rewrite the operator $A_2[\xi]$ (defined in (A.8)) as follows

$$A_2[\xi] = \sum_{i=1}^{k} \lambda_i^{\beta} e_i \otimes (\hat{e}_i[\xi] - e_i).$$

Then applying the Fourier expansion $\hat{e}_i[\xi] - e_i = \sum_{q\geq 1}\langle \hat{e}_i[\xi] - e_i, e_q\rangle e_q$, we can expand $A_2[\xi]$ into

$$\begin{aligned}
&\sum_{i=1}^{k}\sum_{q\geq 1} \lambda_i^{\beta}\langle \hat{e}_i[\xi] - e_i, e_q\rangle e_i\otimes e_q\\
=&\sum_{i=1}^{k} \frac{-\lambda_i^{\beta}}{2} \|\hat{e}_i[\xi] - e_i\|^2 e_i\otimes e_i + \sum_{i=1}^{k}\sum_{q\neq i} \lambda_i^{\beta}\langle \hat{e}_i[\xi] - e_i, e_q\rangle e_i\otimes e_q.
\end{aligned}$$

In the second line we have split up the terms for $q = i$ and $q \neq i$ and have employed identity (B.2) for the first term. Proceeding as for $A_1[\xi]$ now yields the (uniform) order of $\mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$ for the term $A_2[\xi]$.

A_3) Finally we turn to $A_3[\xi]$ in (A.9). Again we use the spectral decomposition $\Gamma = \sum_{q\geq 1} \lambda_q e_q \otimes e_q$ to rewrite this term as

$$A_3[\xi] = \sum_{i=1}^{k} \sum_{q\geq 1} \lambda_q^{\beta} \langle \hat{e}_i[\xi] - e_i, e_q \rangle e_q \otimes (\hat{e}_i[\xi] - e_i).$$

Employing the Fourier expansion of $\hat{e}_i[\xi] - e_i$ for the right factor of the outer product, gives the further expansion

$$\sum_{i=1}^{k} \sum_{q\geq 1} \sum_{l\geq 1} \lambda_q^{\beta} \langle \hat{e}_i[\xi] - e_i, e_q \rangle \langle \hat{e}_i[\xi] - e_i, e_l \rangle e_q \otimes e_l.$$

Though superficially more complicated, this expansion in terms of the product basis $\{e_q \otimes e_l\}_{q,l\in\mathbb{N}}$ (see Section 2.1) is handy, to decompose $A_3[\xi]$ into more simple parts. More precisely we set

$$A_3[\xi] = \sum_{m=1}^{4} A_{3,m}[\xi],$$

where

$$\begin{aligned}
A_{3,1}[\xi] &:= \sum_{i=1}^{k} \lambda_i^{\beta} \langle \hat{e}_i[\xi] - e_i, e_i \rangle^2 e_i \otimes e_i, \\
A_{3,2}[\xi] &:= \sum_{i=1}^{k} \sum_{q\neq i} \lambda_q^{\beta} \langle \hat{e}_i[\xi] - e_i, e_i \rangle \langle \hat{e}_i[\xi] - e_i, e_q \rangle e_q \otimes e_i \\
A_{3,3}[\xi] &:= \sum_{i=1}^{k} \sum_{q\neq i} \lambda_i^{\beta} \langle \hat{e}_i[\xi] - e_i, e_i \rangle \langle \hat{e}_i[\xi] - e_i, e_q \rangle e_i \otimes e_q, \\
A_{3,4}[\xi] &:= \sum_{i=1}^{k} \sum_{l,q\neq i} \lambda_q^{\beta} \langle \hat{e}_i[\xi] - e_i, e_q \rangle \langle \hat{e}_i[\xi] - e_i, e_l \rangle e_q \otimes e_l
\end{aligned}$$

We can now prove the uniform rate of $\mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$ for each of these terms individually.

$\boxed{A_{3,1}:}$ Identity (B.2) in Appendix B implies that $\langle \hat{e}_i[\xi] - e_i, e_i \rangle^2 = \|\hat{e}_i[\xi] - e_i\|^4/4$. Now notice that $A_{3,1}[\xi]$ is a positive definite, diagonal operator, such that its operator norm equals its largest diagonal value, which implies

$$\sup_{\xi\in[1/2,1]} |||A_{3,1}[\xi]|||_{\mathcal{L}} = \sup_{\xi\in[1/2,1]} \max_{i\leq k} \frac{\lambda_i^{\beta}}{4} \|\hat{e}_i[\xi] - e_i\|^4 .$$

As before (in the analysis of $A_{1,1}$) we can use the inequality (B.3) and Lemma A.7, to show that the right side is of order $\mathcal{O}_{\mathbb{P}}(k^{3(\gamma+1)}/N^2) = \mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$.

We bound the Hilbert–Schmidt norm of the remaining terms (since $|||\cdot|||_{\mathcal{L}} \leq |||\cdot|||$) starting with the two middle ones.

$\boxed{A_{3,2}:}$ We begin by noticing that

$$|||A_{3,2}[\xi]|||^2 = \sum_{i=1}^{k} \sum_{q\neq i} \lambda_q^{2\beta} \frac{\|\hat{e}_i[\xi] - e_i\|^4}{4} \langle \hat{e}_i[\xi] - e_i, e_q \rangle^2$$

$$\leq 4\sum_{i=1}^{k}\sum_{q\neq i}\lambda_q^{2\beta}\langle\hat{e}_i[\xi]-e_i,e_q\rangle^2.$$

Here we have used that $\|\hat{e}_i[\xi]-e_i\|\leq\|\hat{e}_i[\xi]\|+\|e_i\|=2$. Proceeding as for the term $A_{1,2}[\xi]$, we see that $|||A_{3,2}[\xi]|||$ is uniformly of order $\mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$.

$\boxed{A_{3,3}:}$ The proof runs among exactly the same lines as for $A_{3,2}[\xi]$: We first observe that

$$|||A_{3,3}[\xi]|||^2\leq 4\sum_{i=1}^{k}\sum_{q\neq i}\lambda_i^{2\beta}\langle\hat{e}_i[\xi]-e_i,e_q\rangle^2$$

and subsequently proceed as for $A_{1,2}[\xi]$ to show that $|||A_{3,3}[\xi]|||$ is uniformly of order $\mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$.

$\boxed{A_{3,4}:}$ A standard calculation shows that the Hilbert–Schmidt norm of $A_{3,4}[\xi]$ is bounded as follows

$$\begin{aligned}|||A_{3,4}[\xi]|||=&\Big\{\sum_{q,l=1}\Big(\lambda_q^{\beta}\sum_{i=1,l\neq i\neq q}^{k}\langle\hat{e}_i[\xi]-e_i,e_q\rangle\langle\hat{e}_i[\xi]-e_i,e_l\rangle\Big)^2\Big\}^{1/2}\\ \leq&\Big\{\Big(\sum_{i=1}^{k}\sum_{q\neq i}\lambda_q^{2\beta}\langle\hat{e}_i[\xi]-e_i,e_q\rangle^2\Big)\Big(\sum_{i=1}^{k}\sum_{l\neq i}\langle\hat{e}_i[\xi]-e_i,e_l\rangle^2\Big)\Big\}^{1/2},\end{aligned}$$

where we have applied Cauchy–Schwarz to the inner part. Next, bounding each factor by the same arguments as in the discussion of $A_{1,2}$ (see (A.13)), we have

$$|||A_{3,4}[\xi]|||\leq\Big(\max_{i\leq k,q\neq i}\frac{\lambda_q^{\beta}}{|\hat{\lambda}_i[\xi]-\xi\lambda_q|}|||\hat{\Gamma}_N[\xi]-\Gamma|||\Big)\Big(\max_{i\leq k,l\neq i}\frac{1}{|\hat{\lambda}_i[\xi]-\xi\lambda_l|}|||\hat{\Gamma}_N[\xi]-\Gamma|||\Big).\quad(A.14)$$

By Lemma A.7 part *ii*) it follows that

$$\max_{i\leq k,q\neq i}\frac{\lambda_q^{\beta}}{\hat{\lambda}_i[\xi]-\xi\lambda_q}=\mathcal{O}_{\mathbb{P}}(1)$$

and by part *i*) of the same Lemma, that

$$\max_{i\leq k,q\neq i}\frac{1}{\hat{\lambda}_i[\xi]-\xi\lambda_q}=\max_{i\leq k,q\neq i}\frac{1}{\xi(\lambda_i-\lambda_q)}\mathcal{O}_{\mathbb{P}}(1)=\mathcal{O}_{\mathbb{P}}(\sqrt{N}).$$

Here we have used Assumption 3.1(7) in the second step (as the difference $\lambda_i-\lambda_q$ is lower bounded by $\min(\lambda_{i-1}-\lambda_i,\lambda_i-\lambda_{i+1})$). Combining this with $|||\hat{\Gamma}_N[\xi]-\Gamma|||=\mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$ (Theorem A.2), it follows from (A.14), that

$$\sup_{\xi\in[1/2,1]}|||A_{3,4}[\xi]|||=\mathcal{O}_{\mathbb{P}}(1/\sqrt{N}).$$

ii) We now want to upper bound $|||U_N[\xi]\hat{\Gamma}_k^{\dagger}[\xi]|||$. For this purpose consider the following decomposition:

$$|||U_N[\xi]\hat{\Gamma}_k^{\dagger}[\xi]|||\leq|||U_N[\xi]\Gamma^{-\zeta}|||\left(|||\Gamma^{\zeta}(\hat{\Gamma}_k^{\dagger}[\xi]-\Gamma_k^{\dagger}/\xi)|||_{\mathcal{L}}+|||\Gamma^{\zeta}\Gamma_k^{\dagger}/\xi|||_{\mathcal{L}}\right),\quad(A.15)$$

where $\zeta = 1/2 - 1/(2\gamma) - \epsilon$ and $\epsilon > 0$ is a positive number specified later. For the inequality we have used Lemma B.1. We now upper bound the factors on the right side of (A.15), beginning with the norm $|||U_N[\xi]\Gamma^{-\zeta}|||$, of which we show:

$$\sup_{\xi\in[1/2,1]} |||U_N[\xi]\Gamma^{-\zeta}||| = \mathcal{O}_{\mathbb{P}}(k^{\epsilon}/\sqrt{N}). \tag{A.16}$$

For this purpose we employ a result from Moricz et al. (1982) (Theorem 3.1). The theorem is in its original form only formulated for real valued random variables, but the proof can be carried over mutatis mutandis to Hilbert space valued variables. It implies that the inequality

$$\mathbb{E}\sup_{\xi\in[1/2,1]} |||\sqrt{N}U_N[\xi]\Gamma^{-\zeta}|||^2 \leq \tilde{C}k^{2\epsilon} \tag{A.17}$$

for some $\tilde{C}$, depending on ϵ, but independent of N, if

$$\mathbb{E}\left\|\left\| \frac{1}{\sqrt{N}}\sum_{i=L}^{H} \varepsilon_i \otimes X_i\Gamma^{-\zeta}\right\|\right\|^2 \leq C\frac{H-L}{N} \tag{A.18}$$

holds for all $1 \leq L \leq H \leq N$ and some C, independent of L, H and N. To verify (A.18) we observe that

$$\begin{aligned}
&\mathbb{E}\left\|\left\| \frac{1}{\sqrt{N}}\sum_{i=L}^{H} \varepsilon_i \otimes X_i\Gamma^{-\zeta}\right\|\right\|^2 = \frac{1}{N}\sum_{i,j=L}^{H} \mathbb{E}\langle \varepsilon_i \otimes X_i\Gamma^{-\zeta}, \varepsilon_j \otimes X_j\Gamma^{-\zeta}\rangle \\
=&\frac{H-L}{N}\sum_{|h|=0}^{H-L-1}\left(1-\frac{|h|}{H-L}\right)\mathbb{E}\langle \varepsilon_0 \otimes X_0\Gamma^{-\zeta}, \varepsilon_h \otimes X_h\Gamma^{-\zeta}\rangle \\
\leq&\frac{H-L}{N}\sum_{h\in\mathbb{Z}}\sqrt{\phi(h)}\mathbb{E}\|\varepsilon_0 \otimes X_0\Gamma^{-\zeta}\|^2,
\end{aligned}$$

where we have used stationarity for the second equality and $\phi-$mixing for the final inequality (for the covariance inequality for mixing we refer to Dehling (1983) equation (3.17) with $s = r = 2$). By Assumption 3.1(3) the sum $\sum_{h\in\mathbb{Z}}\sqrt{\phi(h)}$ is finite. Thus we only have to prove that $\mathbb{E}\|\varepsilon_0 \otimes X_0\Gamma^{-\zeta}\| < \infty$ to get (A.18). By the Cauchy–Schwarz inequality it suffices to show $\mathbb{E}\|\varepsilon_0\|^4, \mathbb{E}\|X_0\Gamma^{-\zeta}\|^4 < \infty$ separately, where $\mathbb{E}\|\varepsilon_0\|^4 < \infty$ by assumption. For the remaining term note that

$$\begin{aligned}
&\mathbb{E}\|X_0\Gamma^{-\zeta}\|^4 = \mathbb{E}\left\|\sum_{n\in\mathbb{N}}\lambda_n^{-\zeta}\langle X_0, e_n\rangle e_n\right\|^4 = \mathbb{E}\left(\sum_{n\in\mathbb{N}}\lambda_n^{-2\zeta}\langle X_0, e_n\rangle^2\right)^2 \\
=&\sum_{m,n\in\mathbb{N}}\lambda_n^{-2\zeta}\lambda_m^{-2\zeta}\mathbb{E}\langle X_0, e_n\rangle^2\langle X_0, e_m\rangle^2 \leq \sum_{m,n\in\mathbb{N}}\lambda_n^{-2\zeta}\lambda_m^{-2\zeta}\sqrt{\mathbb{E}\langle X_0, e_n\rangle^4\mathbb{E}\langle X_0, e_m\rangle^4} \\
\leq&\sum_{m,n\in\mathbb{N}} C\lambda_n^{-2\zeta}\lambda_m^{-2\zeta}\mathbb{E}\langle X_0, e_n\rangle^2\mathbb{E}\langle X_0, e_m\rangle^2 = C\left(\sum_{n\in\mathbb{N}}\lambda_n^{-2\zeta+1}\right)^2 \leq C\left(\sum_{n\in\mathbb{N}} n^{\gamma(2\zeta-1)}\right)^2.
\end{aligned}$$

The last sum is finite, as by choice of $\zeta = 1/2 - 1/(2\gamma) - \epsilon$ we have $\gamma(2\zeta - 1) < -1$. In the above calculations we have used the Cauchy–Schwarz inequality in the first, Assumption 3.1(4) in

the second and Assumption 3.1(6) in the third inequality. We have hence shown (A.18), which -as we have argued- implies (A.17), which again implies (A.16).

We now bound the second factor in (A.15) analyzing the term $\|\!|\Gamma^\zeta(\hat{\Gamma}_k^\dagger[\xi] - \Gamma_k^\dagger/\xi)|\!\|_{\mathcal{L}}$. Notice that

$$\hat{\Gamma}_k^\dagger[\xi] - \Gamma_k^\dagger/\xi = \sum_{i=1}^k \frac{\hat{e}_i \otimes \hat{e}_i}{\hat{\lambda}_i[\xi]} - \sum_{i=1}^k \frac{e_i \otimes e_i}{\xi\lambda_i} = \sum_{j=1}^4 B_j[\xi], \tag{A.19}$$

where

$$\begin{aligned}
B_1[\xi] &:= \sum_{i=1}^k (\hat{e}_i[\xi] - e_i) \otimes \frac{\hat{e}_i[\xi] - e_i}{\hat{\lambda}_i[\xi]} \\
B_2[\xi] &:= \sum_{i=1}^k (\hat{e}_i[\xi] - e_i) \otimes \frac{e_i}{\hat{\lambda}_i[\xi]} \\
B_3[\xi] &:= \sum_{i=1}^k e_i \otimes \frac{\hat{e}_i[\xi] - e_i}{\hat{\lambda}_i[\xi]} \\
B_4[\xi] &:= \sum_{i=1}^k e_i \otimes e_i \frac{\xi\lambda_i - \hat{\lambda}_i[\xi]}{\hat{\lambda}_i[\xi]\lambda_i\xi}
\end{aligned}$$

In the next step we have to show that

$$\sup_{\xi\in[1/2,1]} \|\!|\Gamma^\zeta B_j[\xi]|\!\|_{\mathcal{L}} = \mathcal{O}_{\mathbb{P}}(k^{(\gamma+1)/2+\gamma\epsilon}) \tag{A.20}$$

for each $1 \le j \le 4$. For the sake of brevity, we only present the proofs for $B_2[\xi]$ and $B_4[\xi]$, as $B_1[\xi]$ and $B_3[\xi]$ can be treated by similar techniques.

$\boxed{B_2\text{:}}$ Using a Fourier expansion of the difference $\hat{e}_i[\xi] - e_i$ gives

$$\Gamma^\zeta(\hat{e}_i[\xi] - e_i) = \sum_{q\ge1} \langle \hat{e}_i[\xi] - e_i, e_q\rangle \Gamma^\zeta e_q = \sum_{q\ge1} \lambda_q^\zeta \langle \hat{e}_i[\xi] - e_i, e_q\rangle e_q.$$

Separating the terms where $i = q$ and $i \neq q$ yields

$$\Gamma^\zeta B_2[\xi] = B_{2,1}[\xi] + B_{2,2}[\xi],$$

where

$$B_{2,1}[\xi] := \sum_{i=1}^k \frac{\lambda_i^\zeta}{\hat{\lambda}_i[\xi]} \langle \hat{e}_i[\xi] - e_i, e_i\rangle e_i \otimes e_i \qquad B_{2,2}[\xi] := \sum_{i=1}^k \sum_{q\neq i} \frac{\lambda_q^\zeta}{\hat{\lambda}_i[\xi]} \langle \hat{e}_i[\xi] - e_i, e_q\rangle e_q \otimes e_i. \tag{A.21}$$

Now, proceeding as for $A_1[\xi]$ we have

$$\begin{aligned}
\|\!|B_{2,1}[\xi]|\!\|_{\mathcal{L}} &= \Big\|\!\Big| \sum_{i=1}^k \frac{\lambda_i^\zeta}{\hat{\lambda}_i[\xi]} \langle \hat{e}_i[\xi] - e_i, e_i\rangle e_i \otimes e_i \Big|\!\Big\|_{\mathcal{L}} \\
&\le \max_{1\le i\le k} \frac{\lambda_i^\zeta}{2\hat{\lambda}_i[\xi]} \|\hat{e}_i[\xi] - e_i\|^2 \le 2 \max_{1\le i\le k} \frac{\lambda_i^\zeta}{\hat{\lambda}_i[\xi]} \frac{\|\!|\hat{\Gamma}_N[\xi] - \xi\Gamma|\!\|_{\mathcal{L}}^2}{\min(\lambda_{i-1} - \lambda_i, \lambda_i - \lambda_{i+1})^2}
\end{aligned}$$

(where we have used the identity (B.2) in the first and the bound (B.3) in the second inequality). By part i) of Lemma A.7 we see that

$$\max_{i\leq k} \frac{\lambda_i^\zeta}{\min(\lambda_{i-1}-\lambda_i, \lambda_i-\lambda_{i+1})^2 2\hat{\lambda}_i[\xi]} = \mathcal{O}_{\mathbb{P}}\left(k^{2(\gamma+1)+\gamma(1-\zeta)}\right).$$

Here we have replaced $\hat{\lambda}_i[\xi]$ by $\xi\lambda_i$ in the denominator of the Lemma and then cancelled λ_i^ζ. Recalling that $|||\hat{\Gamma}_N[\xi]-\xi\Gamma|||_{\mathcal{L}}^2 = \mathcal{O}_{\mathbb{P}}(1/N)$ and using Assumption 3.1(6) shows that

$$\sup_{\xi\in[1/2,1]} |||B_{2,1}[\xi]|||_{\mathcal{L}} = \mathcal{O}_{\mathbb{P}}\left(\frac{k^{2(\gamma+1)+\gamma(1-\zeta)}}{N}\right) = \mathcal{O}_{\mathbb{P}}\left(\frac{k^{\gamma+1+\gamma(1-\zeta)}}{\sqrt{N}}\right).$$

Next we consider $B_{2,2}[\xi]$ (defined in (A.21)). By arguments similar to those used in the discussion of the term $A_{1,2}$ (see (A.13)) we have

$$|||B_{2,2}[\xi]||| \leq \max_{i\leq k, q\neq i} \frac{\lambda_q^\zeta}{|\hat{\lambda}_i[\xi]-\xi\lambda_q|\,\hat{\lambda}_i[\xi]} |||\hat{\Gamma}_N[\xi]-\xi\Gamma|||.$$

Again, part i) of Lemma A.7 can be used to replace the empirical eigenvalue by its population counterpart, which (by Assumption 3.1(6)) shows that

$$\max_{i\leq k, q\neq i} \frac{\lambda_q^\zeta}{|\hat{\lambda}_i[\xi]-\xi\lambda_q|\,\hat{\lambda}_i[\xi]} = \mathcal{O}_{\mathbb{P}}\left(k^{(\gamma+1)+\gamma(1-\zeta)}\right)$$

Thus

$$\sup_{\xi\in[1/2,1]} |||B_{2,2}[\xi]||| = \mathcal{O}_{\mathbb{P}}\left(\frac{k^{(\gamma+1)+\gamma(1-\zeta)}}{\sqrt{N}}\right).$$

Putting the estimates for both terms together we see that $|||B_2[\xi]|||_{\mathcal{L}}$ is uniformly of order

$$\mathcal{O}_{\mathbb{P}}\left(\frac{k^{(\gamma+1)+\gamma(1-\zeta)}}{\sqrt{N}}\right) = \mathcal{O}_{\mathbb{P}}\left(k^{(\gamma+1)/2+\gamma\epsilon}\right), \tag{A.22}$$

where the last equality holds by Assumption 3.1(7) and our choice $\zeta = 1/2 - 1/(2\gamma) - \epsilon$.

$\boxed{B_4:}$ We can upper bound the operator norm of $\Gamma^\zeta B_4[\xi]$ as follows:

$$|||\Gamma^\zeta B_4[\xi]|||_{\mathcal{L}} = \Big|\Big|\Big| \sum_{i=1}^k e_i \otimes e_i \frac{\xi\lambda_i - \hat{\lambda}_i[\xi]}{\hat{\lambda}_i[\xi]\lambda_i^{1-\zeta}\xi} \Big|\Big|\Big|_{\mathcal{L}} \leq \max_{1\leq i\leq k} \Big|\frac{\xi\lambda_i - \hat{\lambda}_i[\xi]}{\hat{\lambda}_i[\xi]\lambda_i^{1-\zeta}\xi}\Big|.$$

Note that

$$\sup_{\xi\in[1/2,1]} |\xi\lambda_i - \hat{\lambda}_i[\xi]| \leq \sup_{\xi\in[1/2,1]} |||\hat{\Gamma}_N[\xi]-\xi\Gamma|||_{\mathcal{L}} = \mathcal{O}_{\mathbb{P}}(1/\sqrt{N}),$$

which gives a bound for the numerator. For the denominator we can replace $\hat{\lambda}_i[\xi]$ by the original $\xi\lambda_i$ (part i) of Lemma A.7), which gives a rate

$$\frac{1}{\hat{\lambda}_i[\xi]\lambda_i^{1-\zeta}\xi} = \mathcal{O}_{\mathbb{P}}\Big(\frac{1}{\lambda_i^{2-\zeta}}\Big) = \mathcal{O}_{\mathbb{P}}\big(k^{\gamma(2-\zeta)}\big).$$

Combining these estimates we see that

$$\sup_{\xi\in[1/2,1]} |||\Gamma^{\zeta} B_4[\xi]|||_{\mathcal{L}} = \mathcal{O}_{\mathbb{P}}(k^{\gamma/2-1/2+\gamma\epsilon}) = \mathcal{O}_{\mathbb{P}}(k^{\gamma/2+1/2+\gamma\epsilon}) \ ,$$

where we have used the definition $\zeta = 1/2 - 1/(2\gamma) - \epsilon$ as well as Assumption 3.1(7). We have now shown that indeed (A.20) holds and therefore (see (A.19))

$$\sup_{\xi\in[1/2,1]} |||\Gamma^{\zeta}(\hat{\Gamma}_k^{\dagger}[\xi] - \Gamma_k^{\dagger}/\xi)|||_{\mathcal{L}} = \mathcal{O}_{\mathbb{P}}\left(k^{\gamma/2+1/2+\gamma\epsilon}\right). \tag{A.23}$$

Finally, we derive an upper bound in the second term on the right of (A.15), noting that

$$\Gamma^{\zeta}\Gamma_k^{\dagger}/\xi = \sum_{i\leq k} \frac{\lambda_i^{\zeta}}{\xi\lambda_i} e_i \otimes e_i,$$

which is positive definite and diagonal. Therefore, we obtain for the operator norm

$$\sup_{\xi\in[1/2,1]} |||\Gamma^{\zeta}\Gamma_k^{\dagger}/\xi|||_{\mathcal{L}} = \sup_{\xi\in[1/2,1]} \max_{1\leq i\leq k} \frac{\lambda_i^{\zeta}}{\xi\lambda_i} = \mathcal{O}(k^{\gamma(1-\zeta)}) = \mathcal{O}(k^{(\gamma+1)/2+\gamma\epsilon}). \tag{A.24}$$

Now combing equations (A.16), (A.23) and (A.24) we find, that for any $\epsilon > 0$

$$\begin{aligned} |||U_N[\xi]\hat{\Gamma}_k^{\dagger}[\xi]||| =& \mathcal{O}_{\mathbb{P}}\left(\frac{k^{\epsilon}}{\sqrt{N}}\right)\left(\mathcal{O}_{\mathbb{P}}\left(\ k^{(\gamma+1)/2+\gamma\epsilon}\right) + \mathcal{O}_{\mathbb{P}}\left(\ k^{(\gamma+1)/2+\gamma\epsilon}\right)\right) \\ =& \mathcal{O}_{\mathbb{P}}(k^{(\gamma+1)/2+(1+\gamma)\epsilon}/\sqrt{N}). \end{aligned}$$

Finally, replacing ϵ by $\epsilon/(1+\gamma)$ proves the assertion *ii)* of Lemma (A.6).

□

A.3 Linearization of the test statistic

In the proof of Theorem 3.5 in Section A.1 we have used the stochastic linearization $T[\xi] = T'[\xi] + o_{\mathbb{P}}(1)$, where the objects $T[\xi]$ and $T'[\xi]$ are defined in (A.5) and (A.6) respectively (T' is the linearization of T). That this replacement is valid is a direct consequence of the subsequent Lemma A.5. Before we state our Lemma, we define the linearization function L, which acts on the space of sequential Hilbert–Schmidt operators $\ell^{\infty}(\mathcal{S}(H_1, H_1))$ (see Definition A.1). For convenience we also define the map $\tilde{L}$, which is used to state the long-run variance in Theorem 4.2.

Definition A.4. *Let L_i be the linear functional acting on the space $\ell^{\infty}(\mathcal{S}(H_1, H_1))$ defined pointwise in $F[\xi]$ as*

$$L_i(F[\xi]) := \sum_{q\neq i} \frac{\lambda_q^{\beta}\lambda_i^{\beta}}{\lambda_i - \lambda_q} \langle F[\xi], e_i \otimes e_q\rangle (e_q \otimes e_i).$$

Therewith we define $L(F[\xi]) := \sum_{i\geq 1} L_i(F[\xi]) + L_i(F[\xi])^*$. Moreover we define the map $\tilde{L}$ as $\tilde{L}(F[\xi]) := (\Gamma^{-1/2} \otimes \Gamma^{-1/2})L(F[\xi])$.

Lemma A.5. *Under the assumptions of Theorem 3.5 it holds that*

i) $\sup_{\xi\in[1/2,1]} \|\!|\xi\Gamma^\beta(\hat{\Pi}_k[\xi]-\Pi_k)\Gamma^\beta - L(\hat{\Gamma}_N[\xi]-\xi\Gamma)|\!|\!|_{\mathcal{L}} = o_{\mathbb{P}}(1/\sqrt{N})$

ii) $\sup_{\xi\in[1/2,1]} |\!|\!|U_N[\xi][\xi\hat{\Gamma}_k^\dagger[\xi]\Gamma^\beta - \Gamma^{\beta-1}]|\!|\!| = o_{\mathbb{P}}(1/\sqrt{N})$.

Proof. We first prove *i)*: Plugging in the definition of the projections gives the following expansion:

$$\Gamma^\beta(\hat{\Pi}_k[\xi]-\Pi_k)\Gamma^\beta = \sum_{i=1}^k \Gamma^\beta\left\{(\hat{e}_i[\xi]\otimes\hat{e}_i[\xi]) - (e_i\otimes e_i)\right\}\Gamma^\beta = D_1[\xi]+D_2[\xi]+D_3[\xi],$$

where

$$\begin{aligned}
D_1[\xi] &:= \sum_{i=1}^k \Gamma^\beta((\hat{e}_i[\xi]-e_i)\otimes e_i)\Gamma^\beta \\
D_2[\xi] &:= \sum_{i=1}^k \Gamma^\beta(e_i\otimes(\hat{e}_i[\xi]-e_i))\Gamma^\beta \\
D_3[\xi] &:= \sum_{i=1}^k \Gamma^\beta((\hat{e}_i[\xi]-e_i)\otimes(\hat{e}_i[\xi]-e_i))\Gamma^\beta.
\end{aligned}$$

The proof now consists of two steps: In the first step we show that

$$\xi D_1[\xi] = \sum_{i\geq 1} L_i(\hat{\Gamma}_N[\xi]-\xi\Gamma). \tag{A.25}$$

As $D_2[\xi] = D_1[\xi]^*$, this implies $\xi(D_1[\xi]+D_2[\xi]) = L(\hat{\Gamma}_N[\xi]-\xi\Gamma)$. In the second step we establish that $D_3[\xi]$ is uniformly of order $o_{\mathbb{P}}(1)$.

Step 1:

Using the diagonal representation $\Gamma^\beta = \sum_{q\geq 1}\lambda_q^\beta e_q\otimes e_q$ and the identity $\Gamma^\beta e_i = \lambda_i^\beta e_i$, we can rewrite $D_1[\xi]$ as follows:

$$D_1[\xi] = \sum_{i=1}^k\sum_{q\neq i}\lambda_q^\beta\lambda_i^\beta\langle\hat{e}_i[\xi]-e_i,e_q\rangle(e_q\otimes e_i) + \sum_{i=1}^k\lambda_i^{2\beta}\langle\hat{e}_i[\xi]-e_i,e_i\rangle(e_i\otimes e_i) =: D_{1,1}[\xi]+D_{1,2}[\xi].$$

Here $D_{1,1}[\xi], D_{1,2}[\xi]$ are defined in the obvious way. We first show that $D_{1,2}[\xi]$ is negligible. For this purpose we use (B.2), to see that

$$|\!|\!|D_{1,2}[\xi]|\!|\!|_{\mathcal{L}} = \Big|\!\Big|\!\Big| \sum_{i=1}^k \frac{-\lambda_i^{2\beta}}{2}\left\|\hat{e}_i[\xi]-e_i\right\|^2(e_i\otimes e_i)\Big|\!\Big|\!\Big|_{\mathcal{L}} \leq \max_{1\leq i\leq k}\frac{\lambda_i^{2\beta}}{2}\left\|\hat{e}_i[\xi]-e_i\right\|^2.$$

The maximum is smaller than a multiple of the right side of (A.12), which is uniformly of order $o_{\mathbb{P}}(1/\sqrt{N})$. Next we turn our attention to $D_{1,1}[\xi]$. Applying identity (B.1), $D_{1,1}[\xi]$ can be rewritten as

$$\sum_{i=1}^k\sum_{q\neq i}\lambda_q^\beta\lambda_i^\beta\frac{\langle\hat{\Gamma}_N[\xi]-\xi\Gamma,\hat{e}_i[\xi]\otimes e_q\rangle}{(\hat{\lambda}_i[\xi]-\xi\lambda_q)}(e_q\otimes e_i).$$

We can now show two things: Firstly that in the above representation we can replace the empirical eigenfunction $\hat{e}_i[\xi]$ and eigenvalue $\hat{\lambda}_i[\xi]$ by their respective population counterparts e_i and $\xi\lambda_i$. Secondly,

we can let the outer sum over i run from 1 to ∞, all of this while incurring only an error of size $o_{\mathbb{P}}(1/\sqrt{N})$. More precisely we get with

$$\tilde{D}_{1,1}[\xi] = \sum_{i=1}^{k} \sum_{q \neq i} \lambda_q^\beta \lambda_i^\beta \frac{\langle \hat{\Gamma}_N[\xi] - \xi\Gamma, e_i \otimes e_q \rangle}{\xi(\lambda_i - \lambda_q)} (e_q \otimes e_i)$$

that

$$\sup_{\xi \in [1/2,1]} |||D_{1,1}[\xi] - \tilde{D}_{1,1}[\xi]|||_{\mathcal{L}} = o_{\mathbb{P}}(1/\sqrt{N}). \tag{A.26}$$

The proof of this fact requires the estimation of some further remainder terms, which we defer to the below Lemma A.6. We have now established that $\tilde{D}_{1,1}[\xi] = D_1[\xi] + o_{\mathbb{P}}(1/\sqrt{N})$ and since $\xi\tilde{D}_{1,1}[\xi] = \sum_{i \geq 1} L_i(\hat{\Gamma}_N[\xi] - \xi\Gamma)$ we have shown (A.25), which concludes the first step of this proof.

Step 2:

Next, we show that $|||D_3[\xi]|||$ is uniformly of order $o_{\mathbb{P}}(1/\sqrt{N})$. Using the Fourier expansion $\hat{e}_i[\xi] - e_i = \sum_{q \geq 1} \langle \hat{e}_i[\xi] - e_i, e_q \rangle e_q$ two times, we can rewrite $D_3[\xi]$ as follows:

$$\begin{aligned} D_3[\xi] &= \sum_{i=1}^{k} \sum_{r,q \geq 1} \langle \hat{e}_i[\xi] - e_i, e_q \rangle \langle \hat{e}_i[\xi] - e_i, e_r \rangle \Gamma^\beta (e_q \otimes e_r) \Gamma^\beta \\ &= \sum_{i=1}^{k} \sum_{r,q \geq 1} \lambda_q^\beta \lambda_r^\beta \langle \hat{e}_i[\xi] - e_i, e_q \rangle \langle \hat{e}_i[\xi] - e_i, e_r \rangle (e_q \otimes e_r) \end{aligned}$$

In the second equality we have used the fact that $\Gamma^\beta e_q = \lambda_q^\beta e_q$. Consequently

$$|||D_3[\xi]||| = \Big\{ \sum_{r,q \geq 1} \Big(\sum_{i=1}^{k} \lambda_q^\beta \lambda_r^\beta \langle e_q, \hat{e}_i[\xi] - e_i \rangle \langle \hat{e}_i[\xi] - e_i, e_r \rangle \Big)^2 \Big\}^{1/2}.$$

By applying the Cauchy–Schwarz inequality to the squared sum, we get

$$|||D_3[\xi]||| \leq \sum_{r \geq 1} \sum_{i=1}^{k} \lambda_r^{2\beta} \langle \hat{e}_i[\xi] - e_i, e_r \rangle^2.$$

The proof that

$$\sup_{\xi \in [1/2,1]} |||D_3[\xi]||| = o_{\mathbb{P}}(1/\sqrt{N})$$

is now conducted by similar techniques as for the term A_1 in the proof of Lemma A.3 and therefore omitted.

Finally we turn to the proof of part *ii*) of this Lemma. Since this proof is technically very similar to part *ii*) of Lemma A.3 we only sketch the idea: We begin by the simple upper bound

$$|||U_N[\xi][\xi\hat{\Gamma}_k^\dagger[\xi]\Gamma^\beta - \Gamma^{\beta-1}]||| \leq |||U_N[\xi]\Gamma^{-\zeta}||| \, |||\Gamma^\zeta[\xi\hat{\Gamma}_k^\dagger[\xi]\Gamma^\beta - \Gamma^{\beta-1}]|||_{\mathcal{L}}.$$

Here $\zeta = 1/2 - 1/(2\gamma) - \epsilon$ and $\epsilon > 0$ is a positive number, which can be chosen arbitrarily small. By (A.16) we know that uniformly $|||U_N[\xi]\Gamma^{-\zeta}||| = \mathcal{O}_{\mathbb{P}}(k^\epsilon/\sqrt{N})$. Thus it suffices to show that the factor

$\|\Gamma^\zeta[\xi\hat{\Gamma}_k^\dagger[\xi]\Gamma^\beta - \Gamma^{\beta-1}/\xi]\|_{\mathcal{L}}$ decays at some arbitrarily small, polynomial speed in N, to get the assertion. We upper bound it by the sum

$$\|\Gamma^\zeta[\xi\hat{\Gamma}_k^\dagger[\xi] - \Gamma_k^\dagger]\Gamma^\beta]\|_{\mathcal{L}} + \|\Gamma^{\beta-1+\zeta}\Pi_k - \Gamma^{\beta-1+\zeta}\|_{\mathcal{L}} =: F_1[\xi] + F_2,$$

where $F_1[\xi], F_2$ are defined in the obvious way, and analyse the terms separately. For $F_1[\xi]$, we use similar techniques as in the proof of Lemma A.3 (after equation (A.19)). Notice that we can indeed show convergence to 0 as additional smoothing is applied (by Γ^β from the right). The proof for F_2 is rather simple: $\Gamma^{\beta-1+\zeta} - \Gamma^{\beta-1+\zeta}\Pi_k$ is symmetric, positive definite and can be expressed (by the spectral theorem in Section 2.1) as $\sum_{q>k} \lambda_q^{\beta-1+\zeta} e_q \otimes e_q$. Thus

$$\|\Gamma^{\beta-1+\zeta}\Pi_k - \Gamma^{\beta-1+\zeta}\|_{\mathcal{L}} = \lambda_k^{\beta-1+\zeta} = \mathcal{O}(k^{-\gamma(\beta-1+\zeta)}).$$

This concludes the proof. □

In the proof of Lemma A.5, we have used the identity (A.26) for the final step in the linearization. In the Lemma below we give three upper bounds, which combined directly imply (A.26).

Lemma A.6. *Under the assumptions of Theorem 3.5 the following identities hold:*

i)

$$\sup_{\xi\in[1/2,1]} \Big\| \sum_{i>k} L_i(\hat{\Gamma}_N[\xi] - \xi\Gamma) \Big\| = o_{\mathbb{P}}(1/\sqrt{N}).$$

ii)

$$\sup_{\xi\in[1/2,1]} \Big\| \sum_{i=1}^{k} \sum_{q\neq i} \lambda_q^\beta \lambda_i^\beta \frac{\langle \hat{\Gamma}_N[\xi] - \xi\Gamma, (\hat{e}_i[\xi] - e_i) \otimes e_q \rangle}{(\hat{\lambda}_i[\xi] - \xi\lambda_q)} (e_q \otimes e_i) \Big\| = o_{\mathbb{P}}(1/\sqrt{N}).$$

iii)

$$\sup_{\xi\in[1/2,1]} \Big\| \sum_{i=1}^{k} \sum_{q\neq i} \lambda_q^\beta \lambda_i^\beta \frac{\langle \hat{\Gamma}_N[\xi] - \xi\Gamma, e_i \otimes e_q \rangle (\xi\lambda_i - \hat{\lambda}_i[\xi])}{(\hat{\lambda}_i[\xi] - \xi\lambda_q)(\xi(\lambda_i - \lambda_q))} (e_q \otimes e_i) \Big\| = o_{\mathbb{P}}(1/\sqrt{N}).$$

Proof.

i) It follows by standard calculations, that

$$\sup_{\xi\in[1/2,1]} \|\sum_{i>k} L_i(\hat{\Gamma}_N[\xi] - \xi\Gamma)\| \leq C_N \sup_{\xi\in[1/2,1]} \|\hat{\Gamma}_N[\xi] - \xi\Gamma\|,$$

with (recall that $k := k(N)$)

$$C_N := \max_{i>k, q\neq i} \frac{\lambda_i^\beta \lambda_q^\beta}{(\lambda_i - \lambda_q)}.$$

Lemma A.7 implies that

$$\max_{q\neq i} \frac{\lambda_q^\beta}{|\lambda_i - \lambda_q|} = \mathcal{O}(1).$$

Hence $C_N = o(1)$, as $N \to \infty$. Since uniformly (in ξ) $\|\hat{\Gamma}_N[\xi] - \xi\Gamma\| = \mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$ (see Theorem A.2), the assertion follows.

ii) The squared Hilbert–Schmidt norm of the term in part *ii*) for some fixed ξ equals

$$\sum_{i=1}^{k}\sum_{q\neq i}\lambda_q^{2\beta}\lambda_i^{2\beta}\frac{\langle\hat{\Gamma}_N[\xi]-\xi\Gamma,(\hat{e}_i[\xi]-e_i)\otimes e_q\rangle^2}{(\hat{\lambda}_i[\xi]-\xi\lambda_q)^2} \tag{A.27}$$

and can be upper bounded by

$$\begin{aligned}&\max_{i\leq k,q\neq i}\frac{\lambda_q^{2\beta}}{(\hat{\lambda}_i[\xi]-\xi\lambda_q)^2}\sum_{i=1}^{k}\lambda_i^{2\beta}\sum_{q\neq i}\langle\hat{\Gamma}_N[\xi]-\xi\Gamma,(\hat{e}_i[\xi]-e_i)\otimes e_q\rangle^2\\ \leq&\mathcal{O}_{\mathbb{P}}(1)\sum_{i=1}^{k}\lambda_i^{2\beta}\|(\hat{\Gamma}_N[\xi]-\xi\Gamma)(\hat{e}_i[\xi]-e_i)\|^2,\end{aligned}$$

where we have used Lemma A.7, part *ii*). We further bound the right factor

$$\begin{aligned}&\sum_{i=1}^{k}\lambda_i^{2\beta}\|(\hat{\Gamma}_N[\xi]-\xi\Gamma)(\hat{e}_i[\xi]-e_i)\|^2\leq C\sum_{i=1}^{k}\frac{\lambda_i^{2\beta}|||\hat{\Gamma}_N[\xi]-\xi\Gamma|||_{\mathcal{L}}^4}{\xi^2\min(\lambda_{i-1}-\lambda_i,\lambda_i-\lambda_{i+1})^2}\\ \leq&Ck|||\hat{\Gamma}_N[\xi]-\xi\Gamma|||_{\mathcal{L}}^4\max_{1\leq i\leq k}\frac{\lambda_i^{2\beta}}{\min(\lambda_{i-1}-\lambda_i,\lambda_i-\lambda_{i+1})^2}\end{aligned}$$

Here we have used inequality (B.3) in the first step and the fact that $1/\xi\leq 2$ in the second. Recall that $|||\hat{\Gamma}_N[\xi]-\xi\Gamma|||_{\mathcal{L}}^4$ is uniformly of order $\mathcal{O}_{\mathbb{P}}(1/N^2)$ (see Theorem A.2) and that the ratio of eigenvalues on the right is bounded by part *iii*) of Lemma A.7. Therefore, the term in (A.27) is of order $\mathcal{O}_{\mathbb{P}}(k/N^2)=o_{\mathbb{P}}(N^{-3/2})$ (by Assumption 3.1(7)).

iii) With similar arguments as in *ii*), one sees that the squared Hilbert–Schmidt norm of the term in *iii*) is bounded by

$$\mathcal{O}_{\mathbb{P}}(1)\sum_{i=1}^{k}|||\hat{\Gamma}_N[\xi]-\xi\Gamma|||^2|\hat{\lambda}_i[\xi]-\xi\lambda_i|^2=\mathcal{O}_{\mathbb{P}}(k/N^2),$$

which is $o_{\mathbb{P}}(N^{-3/2})$ and thus yields the desired result.

□

A.4 Convergence results for empirical eigenvalues

In this section we collect a few results on the convergence speed of the empirical eigenvalues to their population counterparts, which are used at several places in this paper.

Lemma A.7. *Under the assumptions of Theorem 3.5, it holds that*

i)

$$\sup_{\xi\in[1/2,1]}\max_{1\leq i\leq k}|\hat{\lambda}_i[\xi]-\xi\lambda_i|=o_{\mathbb{P}}(k^{-\gamma-1})$$

ii)

$$\max_{1\leq i\leq k}\max_{q\neq i}\left|\frac{\lambda_q^{\beta}}{\hat{\lambda}_i[\xi]-\xi\lambda_q}\right|=\mathcal{O}_{\mathbb{P}}(1)$$

iii)

$$\sup_{\xi\in[1/2,1]} \max_{1\leq i\leq k} \max_{q\neq i} \left| \frac{\lambda_q^\beta}{\lambda_i - \lambda_q} \right| = \mathcal{O}(1)$$

Proof. The proof of *i)* follows by an application of part *iv)* of Lemma B.3 below, together with Theorem A.2: According to the former $|\hat{\lambda}_i[\xi] - \lambda_i\xi| \leq \|\hat{\Gamma}_N[\xi] - \xi\Gamma\|_{\mathcal{L}}$ and according to the latter $\|\hat{\Gamma}_N[\xi] - \xi\Gamma\|_{\mathcal{L}}$ is uniformly (in ξ) of order $\mathcal{O}_{\mathbb{P}}(1/\sqrt{N})$. Recalling Assumption 3.1(7) we note that $1/\sqrt{N} = o_{\mathbb{P}}(k^{-\gamma-1})$, which concludes the proof.

For *ii)* we first notice that according to *i)* we can replace $\hat{\lambda}_i[\xi]$ by is population counterpart $\xi\lambda_i$. Since $1/\xi$ is bounded (by 2), the proof of *ii)* can be reduced to the proof of *iii)*.

We now show *iii)*: Let us define the function

$$f(i,q) := \frac{\lambda_q^\beta}{|\lambda_i - \lambda_q|}.$$

We can make the maximum in *iii)* larger by maximizing f over all $\{(i,q) \in \mathbb{N}^2 : i \neq q\}$. Now the proof works by contradiction: Suppose there was a sequence of $\{(i_n^*, q_n^*)\}_{n\in\mathbb{N}}$, such that $f(i_n^*, q_n^*) \to \infty$. For each tuple the value of the function is finite and thus $i_n^* \to \infty$ or $q_n^* \to \infty$. Now there are three possibilities: $|i_n^*/q_n^*|$ is bounded, goes to 0 or goes to ∞. The case $|i_n^*/q_n^*| \to 0$ can be excluded, as for any fixed i the function $q \mapsto f(i,q)$ is monotonically decreasing in q for $q > i$. Thus if there was a sequence (i_n^*, q_n^*) with $f(i_n^*, q_n^*) \to \infty$ and $|i_n^*/q_n^*| \to 0$ it also holds true for $(i_n^*, i_n^* + 1)$ that $f(i_n^*, i_n^* + 1) \to \infty$, which brings us to the next case of bounded $|i_n^*/q_n^*|$: If $|i_n^*/q_n^*|$ was bounded, it follows directly that

$$\frac{\lambda_{q_n^*}^\beta}{\lambda_{i_n^*} - \lambda_{q_n^*}} = \mathcal{O}\left(\frac{(q_n^*)^{-\beta\gamma}}{(q_n^*)^{-\gamma-1}} \right),$$

which by choice of $\beta > 1+1/\gamma$ is asymptotically vanishing. Finally we consider the case where $|i_n^*/q_n^*| \to \infty$. In this situation

$$\frac{\lambda_{q_n^*}^\beta}{\lambda_{i_n^*} - \lambda_{q_n^*}} = \mathcal{O}\left(\frac{\lambda_{q_n^*}^\beta}{\lambda_{q_n^*}} \right),$$

which is asymptotically vanishing. As a consequence we conclude that

$$\max_{(i,q)\in\mathbb{N}^2 : i\neq q} f(i,q) < \infty,$$

proving the assertion.

□

B Miscellaneous

B.1 Operatortheoretic (in)equalities

We begin with an observation concerning bounds on products of operators.

Lemma B.1. *Suppose three Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$ are given. Let $A \in \mathcal{S}(\mathcal{H}_1, \mathcal{H}_2)$, $B \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_3)$ and $B' \in \mathcal{S}(\mathcal{H}_2, \mathcal{H}_3)$, $A' \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$. Then, it holds that*

$$|||BA||| \leq \|B\|_{\mathcal{L}} |||A||| \quad \text{and} \quad |||B'A'||| \leq |||B'||| |||A'|||_{\mathcal{L}}.$$

In the proof we use the notion of trace for linear operators, which is well defined for *trace class operators* (a class of operators, which in particular includes products of Hilbert–Schmidt operators). For a precise definition of the trace Tr we refer to Section 13.5 in Horváth and Kokoszka (2012).

Proof. We only prove the first inequality, as the proof for the second is identical. For A, B it holds that

$$|||BA|||^2 = \mathrm{Tr}[B^*BAA^*] \leq \mathrm{Tr}[B^*B \ \mathrm{Id} \, |||AA^*|||_{\mathcal{L}}] \leq \mathrm{Tr}[B^*B] |||AA^*|||_{\mathcal{L}} = |||B|||^2 |||A|||_{\mathcal{L}}^2.$$

Here we have used that B^*B, AA^* are positive semi-definite and symmetric. Furthermore, for the first inequality we have used that for symmetric, positive semi-definite operators $O_1, O_2, \tilde{O}_2$ the inequality $\mathrm{Tr}[O_1O_2] \leq \mathrm{Tr}[O_1\tilde{O}_2]$ holds, if $\tilde{O}_2 - O_2$ is positive semi-definite (Loewner order). This result extends directly to Hilbert spaces from the finite dimensional case and is therefore not proven. □

Next we want to discuss in which sense the eigensystems of two similar operators are also similar. For this purpose we have to determine how we deal with the non-uniqueness of eigenfunctions:

Remark B.2. Let A, B be two compact, self-adjoint, positive semi-definite operators with eigenvalues (in non-increasing order) and corresponding eigenfunctions α_j, a_j and β_k, b_k respectively. The eigenfunctions are only determined up to sign, i.e. both a_i and $-a_i$ are eigenfunctions of A, belonging to the i-th eigenvalue α_i. However in order to make a comparison of, say the $i-$th eigenfunctions a_i of A and b_i of B meaningful, we have to consider the minimum $\min(\|a_i - b_i\|, \|a_i + b_i\|)$ (otherwise even "the same" eigenfunctions with opposing signs would result in a difference $\|a_i - b_i\| = \sqrt{2}$). For sake of notational parsimony we always assume that, comparing two eigenfunctions of different operators, the functions have the same sign, in the sense that already $\|a_i - b_i\| = \min(\|a_i - b_i\|, \|a_i + b_i\|)$.

In the next lemma we provide some identities for eigenfunctions and eigenvalues of self-adjoint operators.

Lemma B.3. *Let A, B be two compact, self-adjoint, positive semi-definite operators with (in non-increasing order) and corresponding eigenfunctions α_j, a_j and β_k, b_k respectively. Furthermore suppose that all eigenvalues of A are distinct, i.e. $\alpha_1 > \alpha_2 > \cdots$. Then it holds that*

i) for $j \neq k$ and $\alpha_j - \beta_k \neq 0$:

$$\langle a_j - b_j, b_k \rangle = \frac{\langle A - B, a_j \otimes b_k \rangle}{\alpha_j - \beta_k}, \tag{B.1}$$

ii) for any pair of normalized vectors v, w:

$$\langle w, v - w \rangle = -\frac{1}{2}\|v - w\|^2, \tag{B.2}$$

iii) for all $i \geq 1$:

$$\|b_i - a_i\| \leq 2\sqrt{2}\frac{\|A - B\|_{\mathcal{L}}}{\min\{\alpha_{i-1} - \alpha_i, \alpha_i - \alpha_{i+1}\}}. \tag{B.3}$$

iv) for all $i \geq 1$:

$$\|\beta_i - \alpha_i\| \leq |||B - A|||_{\mathcal{L}}. \tag{B.4}$$

Proof. Identities *i*) and *ii*) are straightforward adaptions of Lemma 1 in Kokoszka and Reimherr (2013) and for *iii*) and *iv*) we refer to Horváth and Kokoszka (2012) (Lemmas 2.2 and 2.3). □