Electronic Journal of Statistics
Vol. 16 (2022) 6232–6266
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2083

Preprocessing noisy functional data: A multivariate perspective*

Siegfried Hörmann†

Institute of Statistics,
Graz University of Technology, Austria
e-mail: shoermann@tugraz.at

Fatima Jammoul

Institute of Software Design and Security,
FH JOANNEUM, Austria
e-mail: fatima.jammoul@fh-joanneum.at

Abstract: We consider functional data which are measured on a discrete set of observation points. Often such data are measured with additional noise. We explore in this paper the factor structure underlying this type of data. We show that the latent signal can be attributed to the common components of a corresponding factor model and can be estimated accordingly, by borrowing methods from factor model literature. We also show that principal components, which play a key role in functional data analysis, can be accurately estimated by taking such a multivariate instead of a 'functional' perspective. In addition to the estimation problem, we also address testing of the null-hypothesis of iid noise. While this assumption is largely prevailing in the literature, we believe that it is often unrealistic and not supported by a residual analysis.

MSC2020 subject classifications: Primary 62H25, 62R10; secondary 62M10.
Keywords and phrases: Functional data, factor models, high-dimensional statistics, preprocessing, signal-plus-noise.

Received November 2021.

Contents

arXiv: 2012.05824
*Research partly funded by the Austrian Science Fund (FWF) [P 35520]. Research partly funded by the Federal Ministry for Digital and Economic Affairs of the Republic of Austria through the COIN project FIT4BA.
†Corresponding author.

1. Introduction

Functional data analysis (FDA) is concerned with the analysis of data that can naturally be described as curves. In mathematical terms, data are modeled as random curves $(X(s)\colon s \in \mathcal{S})$, where $\mathcal{S}$ is some continuum. Examples where such data arise are very diverse, ranging from high frequency asset price curves over growth curves or pollution level curves, to 2D satellite images or fMRI scans. For a simple presentation we assume without loss of generality that $\mathcal{S} = [0, 1]$. With technological advances, recording and storing this type of data becomes more and more common and hence the corresponding FDA literature has seen a big upsurge over the past years. For an introduction to the topic we refer, for example, to the textbooks of Ramsay and Silverman [44], Ferraty and Vieu [18], [31] or Kokoszka and Reimherr [34].

In practice functional data are not fully observed, but sampled on a discrete set of time points. Consider functional observations $(X_t(s)\colon 0 \le s \le 1)$, $t \ge 1$, and assume we have measurements of it at time points $0 \le s_1 < s_2 < \ldots < s_p \le 1$. A very common additional working hypothesis in FDA literature is that these measurements come with an additional error, so that we actually observe

$$Y_t = (X_t(s_1), \ldots, X_t(s_p))' + (U_{t1}, \ldots, U_{tp})' =: X_t(\boldsymbol{s}) + U_t. \tag{1.1}$$

Henceforth we are going to write $\boldsymbol{s}$ for $(s_1, \ldots, s_p)$ and use the convention that $g(\boldsymbol{s})$ denotes $(g(s_1), \ldots, g(s_p))'$. The errors U_t can, for example, be related to measurement errors. In this paper we focus on the setting where all data are observed at the same time points s_i. This is typically the case for machine recorded data. The goal then is to separate the errors from $X(\boldsymbol{s})$. Most papers (including those cited below) assume that the components $(U_{ti}\colon 1 \le i \le p)$ are iid with zero mean and variance $\sigma_U^2 > 0$ and that $X_t(s)$ and U_t are independent. To recover the full curve $X_t(s)$ or the discretisation $X_t(\boldsymbol{s})$ (henceforth we refer to both objects as the *signal*), a variety of fitting techniques exist. The goal is to acquire an estimate $\hat{X}_t(s)$ or $\hat{X}_t(\boldsymbol{s})$ that is close to the true latent signal. A very common

technique is the basis expansion approach, explained thoroughly in Ramsay and Silverman [44]. Here, the fitted curve is a linear combination of suitable basis functions. Most popular choices are the Fourier basis or B-splines. By adding a roughness penalty to the least squares criterion the smoothness of the curves can be controlled. Other common approaches employ local polynomial regression or kernel smoothing. Notable publications (without claim of completeness) include Cleveland [13], Müller [36], Härdle et al. [28], Wand and Jones [49], Hall and Opsomer [26], Claeskens et al. [12] and Wood [51]. All these methods are based on a *curve-by-curve* principle, i.e., each curve is fit separately, ignoring the rest of the sample. In contrast to this, Staniswalis and Lee [46] have proposed an approach which takes the entire sample into account. The key idea is to estimate the covariance kernel of the X_t's by some smoothing method, and then expand the curve along the obtained (smooth) functional principal components. A variant of this approach is the well known PACE algorithm established in Yao et al. [52]. Rubín and Panaretos [45] focus on the estimation of latent curves from sparsely sampled functional time series.

A common feature of many of these discussed approaches is a smoothing step at some point during the procedure. The degree of smoothness of the latent curves, which is needed to choose the appropriate number of basis functions or the bandwidth of a kernel smoother, is typically unknown and then the result of the analysis is influenced by a non-verifiable working hypothesis. Cross-validation (CV) may look like an attractive route, since the parameter choices then become data driven. To illustrate that the problem is still challenging, we look at the synthetic example in Figure 1. The two rows in the graph illustrate realisations from two random samples (two observations each). The red marks represent the raw data Y_t and the solid lines the underlying signals $X_t(s)$. In the first example (top row) the data generating process (DGP) is such that curves are smooth, with one outlying measurement in the second curve. In the second example (bottom row) the curves possess a non-smooth segment. In order to distinguish between measurement errors or some systematic structure in the signal, we will typically need a higher measurement frequency, i.e., an increase of p. Thus, the accuracy in recovering the signal is tied to its smoothness, and the relevant question is whether p is large enough relative to the degree of smoothness to sufficiently justify a certain approach.

In this paper we want to complement the existing literature on preprocessing discretely sampled functional data by the following topics:

(A) We propagate a method, which is not tied to the smoothness of the underlying curves.

(B) We develop tools for the analysis of the resulting model residuals in (1.1)—a topic which is, to the best of our knowledge, widely ignored in existing contributions.

In context of **(A)**, we analyse the problem from a multivariate perspective. We show in Section 2 that functional data sampled as in (1.1) follow some *factor model* [see, e.g., 35]. The factor model is able to accommodate the functional nature of X_t without requiring smooth curves. The signal underlying the discre-

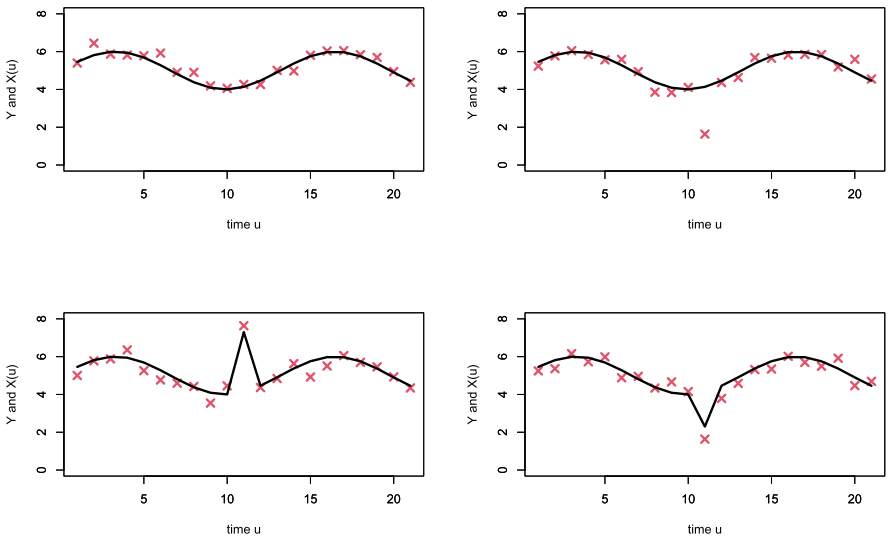

FIG 1. *In this illustrative example, the red marks correspond to measurements Y and the solid lines to the underlying signal $X(s)$.*

tised observations is related to the *common components* of the factor model and thus a natural strategy is to estimate these common components. The two most common techniques for this purpose are either likelihood or PCA based. The PCA technique was shown in our companion paper Hörmann and Jammoul [29] to lead to consistent estimation of the signal under fairly mild assumptions. Here we complement the theoretical results in Hörmann and Jammoul [29] by extending the estimators for $X_t(\boldsymbol{s})$ to consistent estimators for the full curve $X_t(s)$ (Section 2.2) and its functional principal components (Section 2.3). These extensions do not rely on smoothing steps either.

With regards to **(B)**, we note that most papers impose iid error components $(U_{ti}\colon 1 \leq i \leq p)$. However, the interpretation of the errors (aside from measurement errors) is broadly ignored and a thorough residual analysis, which is required for corresponding model diagnostics, is barely addressed in existing contributions. We devote Section 3 to adequate diagnostic tools and explain how the factor model approach leads to a sensible interpretation of the model errors, going beyond measurement errors. In Section 3.1 we develop a frequency domain based test statistic for iid errors and derive its asymptotic null-distribution by exploring the double-asymptotics $p, T \to \infty$. The test statistic in turn also blazes a trail to an empirical approach to determine the number of factors of the underlying factor model via an alternative version of the scree plot. This will be discussed in some detail in Section 3.2.

The rest of the paper provides comprehensive numerical studies. In the simulation studies in Section 4 we consider not only smooth signals (Section 4.1) but also signals which contain discontinuities (Section 4.2). After blurring those

signals with noise, we investigate how well our propagated approach and competing methods are able to recover the signals. Our test statistic for iid noise is evaluated in Section 4.3. Finally, in Section 5 we illustrate our method on some real data examples.

2. Factor model representation

We consider a set of functional data $X_1, \dots, X_T$ defined on a common probability space. Throughout the paper we assume that observations are iid or form a general stationary functional process. The curves $(X_t(s)\colon s \in [0,1])$ are square integrable on $[0,1]$, and hence can be expanded along a sequence of orthogonal basis functions $\{b_k(s)\colon k \geq 1\}$, e.g. the Fourier basis. Then we have $X_t(s) = \sum_{k\geq 1}\langle X_t, b_k\rangle b_k(s)$, where $\langle a, b\rangle = \int_0^1 a(v)b(v)dv$. The convergence is in general only in L^2 sense, but under mild regularity conditions on path properties of X_t we can also obtain pointwise or even uniform convergence. In particular, if the covariance kernel $\Gamma^X(s,s') := \mathrm{Cov}(X_t(s), X_t(s'))$ is continuous and we set $b_k = \varphi_k$, which denote the eigenfunctions of $\Gamma^X(s,s')$, then we obtain as a consequence of Mercer's theorem [see, e.g., 23], that

$$\sup_{s\in[0,1]} E\left|X_t(s) - \mu(s) - \sum_{\ell=1}^{L} x_{t\ell}\varphi_\ell(s)\right|^2 \to 0, \quad L \to \infty, \tag{2.1}$$

where $x_{t\ell} = \int_0^1 (X_t(s) - \mu(s))\varphi_\ell(s)ds$. The functions φ_k are the so-called functional principal components, and define an optimal orthogonal basis system, in the sense of minimising the mean square error

$$\int_0^1 E\left|X_t(s) - \mu(s) - \sum_{\ell=1}^{L}\langle X_t, b_k\rangle b_k(s)\right|^2 ds$$

with respect to the basis functions (b_k). In typical applications the approximation error is already very close to zero with small L (say $L = 5$) or at most moderately sized values of L (say $L = 20$), so that assuming a finite dimensional representation

$$X_t(s) = \sum_{k=1}^{L}\langle X_t, b_k\rangle b_k(s), \quad \text{for some } L \geq 1 \tag{2.2}$$

is no more than a theoretical restriction, which imposes no practical limitation of generality, if L is allowed to be chosen large enough.

A basic requirement for our proposed method is that all curves are sampled at the same time points $0 \leq s_1 < s_2 < \ldots < s_p \leq 1$. This is a very common setting for machine recorded data. We note that sampling points need not be equidistant though. We will assume throughout a general signal-plus-noise structure as in (1.1).

The following representation theorem for functional data observed as in (1.1) holds.

Proposition 2.1. *Suppose* (1.1) *and* (2.2) *hold. Let* U_t *be independent of* X_t *and assume that* $EU_t = 0$ *and* $\mathrm{Var}(U_t)$ *is diagonal. Then* Y_t *follows an* L*-factor model.*

Proof. We show that there exists a matrix $B \in \mathbb{R}^{p\times L}$ such that

$$Y_t = \mu(\boldsymbol{s}) + BF_t + U_t, \tag{2.3}$$

where $\mu(s) = EX_t(s)$, $EF_t = 0$, $\mathrm{Var}(F_t) = I_L$ (the identity matrix in $\mathbb{R}^L$) and $\mathrm{Cov}(F_t, U_t) = 0$. We note that by the imposed stationarity the covariance kernel $\Gamma^X(s, s')$ does not depend on t. Using (2.2) the Karhunen-Loève expansion gives

$$X_t(s) = \mu(s) + \sum_{\ell=1}^{L} x_{t\ell}\varphi_\ell(s), \tag{2.4}$$

where $\varphi_\ell(s)$ are the eigenfunctions of the covariance operator Γ^X. The scores $(x_{t\ell}\colon \ell \geq 1)$ are uncorrelated and $\mathrm{Var}(x_{t\ell}) = \lambda_\ell$, where λ_ℓ are the eigenvalues of Γ^X (in decreasing order). See, e.g., Bosq [7] for details. Define $B := (\sqrt{\lambda_1}\varphi_1(\boldsymbol{s}), \ldots, \sqrt{\lambda_L}\varphi_L(\boldsymbol{s}))$. Moreover, define

$$F_t = (x_{t1}/\sqrt{\lambda_1}, \ldots, x_{tL}/\sqrt{\lambda_L})'.$$

This yields the desired representation. □

In factor model language BF_t are called the *common components* of Y_t and our problem is reduced to the estimation of these common components. For this purpose we can resort to a rich literature, especially from macroeconomics, where factor models are used to model markets with many assets. See, e.g., Stock and Watson [47, 48] and Forni and Lippi [19]. In this context, Chamberlain and Rothschild [10] have shown that it is useful to allow also for a certain degree of dependence in the *idiosyncratic noise components* U_{ti}. This setting then refers to *approximate factor models*. Some of the features employed in econometric applications are natural and useful in our context, too: (1) The dimension p of our sampling points $\boldsymbol{s}$ is large and allowed to diverge with increasing sample size. (2) The functional data X_t may be time-dependent, i.e., form a functional time series. (3) In a realistic framework, the errors $(U_{ti}\colon 1 \leq i \leq p)$ in (1.1) might be correlated at small lags.

Next to conceptual papers proposing different variants of factor models, there is also a profound literature on estimation theory for these models. In particular we refer to the papers of Bai [1], Bai and Li [2], Choi [11], Fan et al. [16], Bai and Liao [4] and Bai and Li [3]. In context of dynamic factor models we refer to Forni et al. [20, 21]. How these methods may be used in the current context will be discussed in the next section.

We conclude here with two important remarks.

Remark 1. *It is common in FDA to smooth data, even if by their very nature they come without relevant measurement errors (e.g., annual temperature curves generated from daily data, intraday stock prices, etc.). In this case it needs to be*

clarified how the residual noise is to be interpreted. The translation of our problem into factor model language gives a mathematical/statistical meaning to the noise U_t which goes beyond measurement errors. The U_t define the idiosyncratic components *of Y_t, which are characterised by being uncorrelated or, more generally, being weakly correlated in a certain sense to be specified. The components of U_t represent "unsystematic" fluctuations in our functional trajectories.*

Remark 2. *We consider the representation/approximation of $X_t(s)$ via (2.4) as the essence of the functional nature of the data. It is manifested by the "co-movement" of $X_t(s)$ and $X_t(s')$ via linear combinations with a limited number of basis functions at abscissae s and s'. The common component in a factor model is the multivariate analogue of this. Specific smoothness conditions on the $\varphi_\ell(s)$ are not relevant for such an interpretation.*

2.1. Estimation approach

As mentioned above, the signal $X_t(\boldsymbol{s})$ is related to the common components of Y_t. The core idea of the algorithm that we pursue is simple and can be summarised as follows:

Core algorithm:

1. Estimate $\mu(\boldsymbol{s})$ by $\hat{\mu}(\boldsymbol{s}) = \frac{1}{T}(Y_1 + \cdots + Y_T)$.
2. Center the data by $\hat{\mu}(\boldsymbol{s})$.
3. Choose an appropriate order L.
4. Approximate $X_t(\boldsymbol{s}) - \mu(\boldsymbol{s})$ through the estimated *common components*: $\hat{B}\hat{F}_t$.
5. Set $\hat{X}_t(\boldsymbol{s}) = \hat{\mu}(\boldsymbol{s}) + \hat{B}\hat{F}_t$.

Steps 3. and 4. can be carried out by many existing approaches for factor models. Bai and Ng [5] is a key reference for determining the dimension L. Hallin and Liška [27] expanded the approach to dynamic factor models. Onatski [38] proposes an approach that uses the empirical distribution (ED) of the eigenvalues of the sample covariance matrix. Owen and Wang [39] use a Bi-Cross-Validation (BCV) technique to estimate the number of factors. Contrary to other approaches, Owen and Wang [39] are not specifically interested in recovering the true number of factors, but rather the number of factors best-suited to recover the underlying signal. In the process of our empirical work, we have investigated the behaviour of the above mentioned estimators. We found that the BCV and ED approaches work best in our FDA context. In Section 3.2 we will propose an empirical method to choose L.

Once L is fixed, there are two main approaches for factor model estimation. One strategy is to utilize principal component analysis, e.g., Chamberlain and Rothschild [10] use this method. PCA is particularly simple to implement and does not require numerically intense stochastic optimization methods. Bai [1] investigated the asymptotic behaviour of both the factors as well as the factor

loadings and—under technical conditions—proved consistency as well as robustness to mild correlation in the error terms.

The second popular strand is based on maximum likelihood. Choi [11] expanded upon previous ideas by describing an efficient estimation for factor models, where the conditional distribution of $U_t|F_1, \dots F_T$ is assumed to be normal with a covariance matrix that is not necessarily diagonal. Bai and Li [2, 3] provide a method involving a quasi-maximum-likelihood approach.

Let us discuss the PCA approach in detail, which can be motivated as follows. Let $Y = (Y_1, \dots, Y_T)$ and define $U = (U_1, \dots, U_T)$ and $F' = (F_1, \dots, F_T)$. Then, assuming zero mean, we can write our model equation (2.3) in the compact matrix form

$$Y = BF' + U. \tag{2.5}$$

In this notation, the objective is to estimate BF' through some estimator $\hat{B}\hat{F}'$. Suppose that F is already known. Then $Y^j = Fb_j + U^j$, which leads to the common least-squares estimator $\hat{b}_j^{\mathrm{LS}} = (F'F)^{-1}F'Y^j$. Here b_j' is the j-th row of B and Y^j and U^j denote the j-th column of Y' and U', respectively. If our data are independent (or satisfy some appropriate weak dependence condition), it holds by the law of large numbers and orthogonality of principal components scores that

$$\frac{1}{T}F'F \xrightarrow{P} I_L \quad (T \to \infty). \tag{2.6}$$

This motivates $\hat{B}_{|F} := \frac{1}{T}YF$ as estimator for B conditional on F. For F in turn we use the empirical principal components and set $\hat{F} = \sqrt{T}\hat{E}$, where $\hat{E} = (\hat{e}_1, \dots, \hat{e}_L)$ are the eigenvectors of $\frac{1}{T}Y'Y$ $(T \times T)$ associated to the L largest eigenvalues $\hat{\gamma}_1 \geq \dots \geq \hat{\gamma}_L$. Then $\frac{1}{T}\hat{F}'\hat{F} = I_L$. In summary $\hat{F} = \sqrt{T}\hat{E}$ and $\hat{B} = \frac{1}{T}Y\hat{F}$, which implies that

$$(\hat{X}_1(\boldsymbol{s}), \dots, \hat{X}_T(\boldsymbol{s})) = \widehat{BF'} := \hat{B}\hat{F}' = Y\hat{E}\hat{E}'. \tag{2.7}$$

We have analysed this estimator for the signal in Hörmann and Jammoul [29] and have shown that under mild technical conditions (see Assumptions 2-4 in the Appendix) this estimator converges uniformly, i.e.,

$$\sup_{1\leq t\leq T} \sup_{1\leq i\leq p} |X_t(s_i) - \hat{X}_t(s_i)| \to 0 \quad (p, T \to \infty)$$

in probability and explicit convergence rates can be obtained. In applications the user is free to choose any estimation method that leads to satisfactory and plausible results. (See Section 3.)

Remark 3. *A well known problem in factor model theory is that factor loadings and the factor scores are not unique. If $O \in \mathbb{R}^{L\times L}$ is some orthogonal matrix, then $BF = (BO)(O'F)$, and $\mathrm{Var}(O'F) = I_L$. This identification issue is not a problem here, because we are primarily interested in the common components BF, which remain well identified.*

2.2. Estimating the full curve

The factor approach does not return a full curve, but an estimate of the noise-free curves at the points $0 \leq s_1 < s_2 < \ldots < s_p \leq 1$. If the goal is to work with full curves, then it is up to the experimenter to choose a discrete-to-function transformation which is designated for noise-free data. The simplest approach, namely linear interpolation, will be considered in this section.

Theorem 2.1. *We consider a sample $Y_1, \ldots, Y_T$ of discretely observed functional data as in (1.1). Let $\hat{X}_t(s_i)$ be the PCA based factor model estimates for the underlying signal at the points $0 = s_1 < s_2 < \ldots < s_p = 1$ as defined in (2.7) and let $\hat{X}_t(s)$, $s \in [0,1]$, be the linear interpolation of these estimates $\hat{X}_t(s_i)$. Denote $\delta = \max_{1 \leq i \leq p-1} |s_{i+1} - s_i|$. Assume that for some $\alpha \in (0,1]$ we have a random variable M_t such that*

$$|X_t(s) - X_t(u)| \leq M_t |s-u|^\alpha \tag{2.8}$$

holds, where $EM_t = m < \infty$. Then under Assumptions 2–4 in the Appendix we have

$$\sup_{s \in [0,1]} |X_t(s) - \hat{X}_t(s)| = O_P\left(\sqrt{\frac{\log p}{T}} + \frac{1}{\sqrt{p}} + \delta^\alpha\right). \tag{2.9}$$

Remark 4. *In the formulation of this theorem it is assumed that L is fixed and known. Like in [29] the result can be extended to the cases where L is replaced by a consistent estimator. It is also possible to derive variants of this theorem where L is allowed to diverge with the sample size T. This, however, requires further technical assumptions. We refer to Theorem 2 in [29].*

In order to extend our results to the full sample paths we require the Lipschitz condition (2.8), which has, for example, been previously considered in Bosq [7, p.169] or in Kallenberg [32]. Prominent examples of processes that fulfill this property include the Brownian and fractional Brownian Motion, hence also processes which are by no means smooth. Note that under these assumptions, the observation points need not be equidistant in order to control size of the modulus of continuity, but merely the largest distance between two knots needs to become small. It is natural to assume that $\delta = O(p^{-1})$ holds, implying that the term δ^α is negligible if $\alpha \geq 1/2$.

2.3. Estimation of eigenfunctions

Functional principal components take a central role in FDA literature [see, e.g., 44]. When data are fully observed, the estimation theory is well established [33, 14, 25]. When data are discretely observed and subject to measurement errors, then obviously estimation theory has to be adapted. The most common strategy is to first estimate the curves using techniques described in the introduction and then to estimate principal components from the empirical covariance operator of the fitted data. Alternatively, one may use eigenfunctions of the non-parametric

estimates of the covariance kernel as suggested in Staniswalis and Lee [46] or Yao et al. [52].

We now show that functional principal components can be estimated quite well from discretely observed and noisy data. Unlike the procedures mentioned before, this does not involve a smoothing step. Let us begin by noting that Lemma 1 in Hörmann and Jammoul [29] shows that under some mild technical assumptions the ℓ-th eigenvalue λ_ℓ of the covariance kernel $\Gamma^X(s,s') = \mathrm{Cov}(X_t(s), X_t(s'))$ may be consistently estimated by $\hat{\gamma}_\ell^Y/p$, which denotes the p-th fraction of the ℓ-th eigenvalue of $\hat{\Sigma}^Y := T^{-1}YY'$ ($\in \mathbb{R}^{p\times p}$). A similar result has been obtained in Benko et al. [6]. These authors also work with the raw data when estimating the eigenvalues. For estimation of eigenfunctions they do, however, use a smoothing step. To formulate our result, we denote the eigenvectors associated to the eigenvalues $\hat{\gamma}_\ell^Y$ by $\hat{\psi}_\ell^Y$. In order to properly describe the relationship between the function φ_ℓ and the vector $\hat{\psi}_\ell^Y$ we define $\tilde{\varphi}_\ell(s) = \sqrt{p}[\hat{\psi}_\ell^Y]_i$ if $s \in [s_i, s_{i+1})$, where $[v]_i$ denotes the i-th component of a vector v. The step-function $\tilde{\varphi}_\ell$ is the proposed estimator for the eigenfunction φ_ℓ. Note that the scaling ensures that $\|\tilde{\varphi}_\ell\|^2 := \int_0^1 \tilde{\varphi}_\ell^2(s) = 1$.

Remark 5. *Eigenfunctions and eigenvectors are of course uniquely defined only up to the sign. In order to ensure that $\tilde{\varphi}_\ell$ indeed is the estimate for φ_ℓ, we assume that $\langle \varphi_\ell, \tilde{\varphi}_\ell \rangle \geq 0$ holds. To lighten the notation, we henceforth assume in the proof of Theorem 2.2 that the inner product is nonnegative for any pair of eigenfunctions and eigenvectors whose difference is being investigated.*

Theorem 2.2. *Let Assumptions 2 and 3 (a) and (b) hold. Assume that the sampling points s_i are equidistant and that*

$$\sup_{s\in[0,1]} E|X(s+h) - X(s)|^2 = O(h) \quad (h \to 0). \tag{2.10}$$

Then if $\alpha_\ell = \min\{\lambda_\ell - \lambda_{\ell+1}, \lambda_{\ell-1} - \lambda_\ell\} \neq 0$, we have

$$\|\varphi_\ell - \tilde{\varphi}_\ell\| = O_P\left(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{T}}\right), \quad \ell \geq 1.$$

Benko et al. [6] have compared their eigenfunction estimators from discretely observed and noisy data to the empirical eigenfunctions $\hat{\varphi}_\ell$ from fully observed data. They show that the error is of smaller order of magnitude than the error between $\hat{\varphi}_\ell$ and φ_ℓ. Since their result is pointwise in s, it is not directly comparable to our L^2 distance. From a technical point of view both results have advantages and disadvantages. Our result holds under milder smoothness conditions. We merely need Assumption (2.10), while they request second order derivatives with a uniformly bounded fourth order moment. Furthermore, we allow for dependence in both the errors and the observations. Benko et al. [6] focus on the iid setup. On the other hand, they allow for more general errors with 8 moments and do not request a regular sampling design.

3. Model diagnostics

A simple diagnostic tool which may help to discern inadequate signal extraction is the inspection of the covariance of the residuals. Consider the fits $\hat{X}_1(\boldsymbol{s}), \ldots, \hat{X}_T(\boldsymbol{s})$ and denote by $\hat{U}_t = Y_t - \hat{X}_t(\boldsymbol{s})$ the residual vectors. Each residual vector $\hat{U}_t$ defines a time series $\hat{U}_{t1}, \ldots \hat{U}_{tp}$. For example, if $(U_{ti} \colon 1 \le i \le p)$ is assumed to be white noise, then this should be reflected in the empirical autocorrelation functions (acf's)

$$\hat{\gamma}_{\hat{U}_t}(h) = \frac{1}{p} \sum_{i=1}^{p-|h|} (\hat{U}_{t,i+h} - \bar{\hat{U}}_{t\cdot})(\hat{U}_{ti} - \bar{\hat{U}}_{t\cdot}). \tag{3.1}$$

Since we have replicates, we may also conclude that

$$\hat{\Gamma}^{\hat{U}} := \frac{1}{T} \sum_{t=1}^{T} (\hat{U}_t - \bar{\hat{U}})(\hat{U}_t - \bar{\hat{U}})' \approx \mathrm{Var}(U_t), \tag{3.2}$$

where $\bar{\hat{U}}$ is the grand mean of $\hat{U}_1, \ldots, \hat{U}_T$. If there is doubt that the noise components are stationary (e.g., if the homogeneous variance assumption is likely to be violated) analysing $\hat{\Gamma}^{\hat{U}}$ may be preferable over investigating the acf's $\hat{\gamma}_{\hat{U}_t}$. If the residual covariances do not conform with the assumptions on the noise variables (e.g., iid noise), this indicates that either these assumptions were incorrect, or that the transformation from discrete to functional data introduced some bias.

In our real data examples (Section 5) we investigate daily mean temperatures and corresponding annual temperature curves from Canada. Following Ramsay et al. [42], the daily data were transformed to annual curves using 65 basis functions and a roughness penalty. In Figure 2 we show the acf's (3.1) of the residual vectors of this penalized B-spline approach at a weather station in St. Margaret's Bay, Nova Scotia, in the year 1993. We also show the heat map representing the lefthand side in (3.2). For better visibility, the heat map is restricted to the first 2 months of the year. Details on the data and the implementation will be given in Section 5. At this stage, we want to draw the readers attention to the spurious oscillation in the acf. If the components of the error vectors were iid—as it is commonly assumed—then the acf should be zero for all lags $\neq 0$. In a slightly more realistic setting we would expect some moderate positive correlation of the errors, which tapers to zero with increasing lag.

3.1. Testing for independent errors

In FDA literature the iid assumption for the error components $U_{t1}, \ldots, U_{tp}$ is strongly prevailing. Below we refer to this assumption as the null hypothesis $\mathcal{H}_0$. Surprisingly, however, on real data this assumption is typically used without providing empirical evidence. To the best of our knowledge, no specific statistical tests have been developed for this problem. Of course, a straight forward strategy

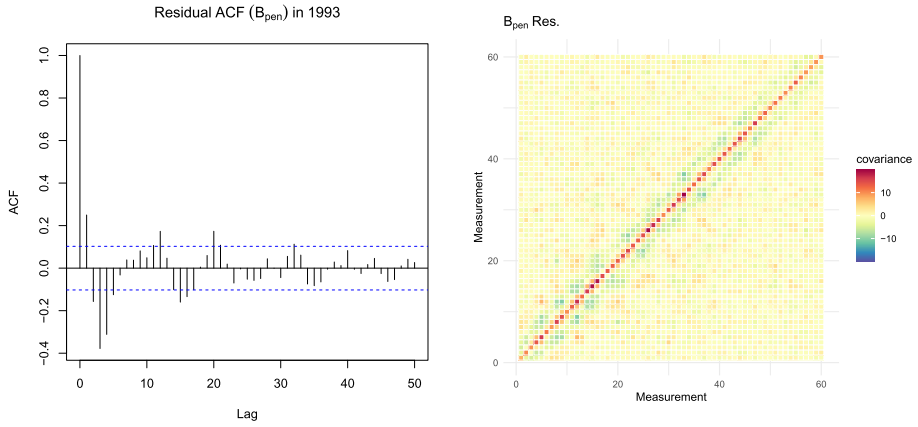

FIG 2. *Autocorrelation function of residual vectors for Penalized B-Splines in the St. Margaret's Bay Data example for the year 1993 and the heat map corresponding to the lefthand side in* (3.2).

is to employ some of the existing white noise tests individually to each residual vector and then to aggregate the information from the resulting T tests. Below we propose a tailor-made test statistic for our setting. To this end we introduce some further notation. We assume throughout that the error vectors U_t, $1 \le t \le T$, are iid. The components will also be iid or stationary, depending on whether we operate under the null hypothesis or the alternative hypothesis.

For some generic random vector $Z = (Z_1, \dots, Z_p)'$ we denote the empirical variance of the components of Z by $S_Z^2 = \frac{1}{p-1}\sum_{k=1}^{p}(Z_k - \bar{Z})^2$. The periodogram is defined as

$$I_Z(\theta) = \frac{1}{p}\left|\sum_{k=1}^{p} Z_k e^{-\mathrm{i}k\theta}\right|^2.$$

Here $\mathrm{i} = \sqrt{-1}$ and $|z|$ is the modulus of a complex number z. We refer to the frequencies $\theta_\ell = \frac{2\pi\ell}{q}$, $1 \le \ell \le q := \lfloor p/2 \rfloor$ as the fundamental frequencies.

Now we choose a subset of fundamental frequencies $\boldsymbol{\theta} = \{\theta_\ell,\ \ell \in \mathcal{F} \subset \{1, \dots, q\}\}$ and denote $f := |\mathcal{F}|$. We allow $\mathcal{F}$ (and hence f) and p to depend on T and our asymptotic statements below are then for $T \to \infty$. Set $\xi = T^{-1}\sum_{t=1}^{T} I_{U_t}(\boldsymbol{\theta})$ and note that ξ is an estimator of the spectral density of the U_t at the fundamental frequencies contained in $\boldsymbol{\theta}$. If the components U_{ti} are iid, then the spectral density is constant and the components of ξ will be roughly constant as well. Our test statistic is thus based on the empirical variance S_ξ^2, which under $\mathcal{H}_0$ shall be accordingly small. Proposition 3.1 below establishes the essential asymptotic result related to the proposed test under the null.

Proposition 3.1. *Assume that $\mathcal{H}_0$ holds and that $EU_{11} = 0$, $EU_{11}^2 = \sigma^2$ and $EU_{11}^4 < \infty$. Let $\hat{\sigma}^2$ denote a consistent estimator of the variance. Then*

$$\Lambda_{\text{fin}} := (f-1)TS_\xi^2/\hat{\sigma}^4 \xrightarrow{\text{d}} \chi_{f-1}^2.$$

If additionally $EU_{11}^8 < \infty$ *and* $f \to \infty$*, and* $f/T \to 0$*, then*

$$\Lambda_{\text{inf}} := \left(TS_\xi^2/\hat{\sigma}^4 - 1\right)\sqrt{(f-1)/2} \xrightarrow{\text{d}} N(0,1).$$

If p is of the same order or of a bigger order of magnitude than T (e.g., this is the case in our real data in Section 5), then taking $\mathcal{F} = \{1, \dots, q\}$ is not theoretically justified by the proposition, since then $f \approx p/2$ and hence $f/T \not\to 0$. We can overcome this problem by thinning out the frequencies $\mathcal{F}$, i.e., we choose some large enough m and only take every m-th frequency. Then $f \approx \frac{p}{2m}$.

In our next result we want to show the proposed test is consistent under the following alternative:

Assumption 1. *[Alternative Hyptothesis] We assume that the process* $U := \{U_{ti} : i \geq 1\}$ *is stationary with absolutely summable autocovariance* γ_U *function and spectral density*

$$g(\theta) := \sum_{h \in \mathbb{Z}} \gamma_U(h) e^{ih\theta}.$$

Additionally we assume that $Var(I_{U_1}(\theta_\ell))$ *is uniformly bounded for all* $\ell \in \mathcal{F}$ *and all dimensions* p*. Finally, denoting* $\bar{g} = \frac{1}{f}\sum_{\ell \in \mathcal{F}} g_U(\theta_\ell)$ *we assume that there is some* $\delta > 0$ *such that*

$$\frac{1}{f-1}\sum_{\ell \in \mathcal{F}} (g(\theta_\ell) - \bar{g})^2 > \delta. \tag{3.3}$$

When f *diverges,* (3.3) *should hold uniformly in* f*.*

Besides mild technical moment assumptions (which hold, e.g., for certain linear processes), our basic requirement under the alternative is that the noise is correlated and hence that the spectral density is not constant. In order to detect such a non-constant spectral density, we have to assure that it varies at the frequencies we have incorporated in our test statistic. This is assured by (3.3). Note that the term in (3.3) does not just depend on f but also on the choice of frequencies. If we select the frequencies θ_ℓ on a regular grid and $f \to \infty$, then we can replace our condition by $\int_0^\pi \left(g(\theta) - \int_0^\pi g(s)ds\right)^2 d\theta > \delta$.

Proposition 3.2. *Consider the setting of Proposition 3.1 and assume that* (1) *holds. Let* $0 \leq h_T = o(T)$*. Then* $\Lambda_{\text{fin}}/h_T \to \infty \quad (T \to \infty)$*. If additionally* $f = f(T) \to \infty$*,* $f/T \to 0$ *and if* $0 \leq h_T = o(T\sqrt{f})$ *then* $\Lambda_{\text{inf}}/h_T \to \infty \quad (T \to \infty)$*.*

In practice the U_{ti} are latent and the test will be applied to the residuals $\hat{U}_{ti} = Y_{ti} - \hat{X}_t(s_i)$. This gives then rise to the test statistics $\hat{\Lambda}_{\text{fin}}$ and $\hat{\Lambda}_{\text{inf}}$. The results above do not account for the effect of the estimation error $\delta_{ti} := \hat{U}_{ti} - U_{ti} = \hat{X}_t(s_i) - X_t(s_i)$. We have experienced in simulations that frequencies close to zero seem to get eliminated in the time series $\{\hat{U}_{ti} : 1 \leq i \leq p\}$ and hence we do not accurately estimate the spectral densities at very low frequencies; see Figure 3. One can think of the low frequency part of the noise as a superposition of slowly swinging sinusoids, which seem to be erroneously attributed to the latent signal. Note that even a very small perturbation error, such as $\sup_{1 \leq i \leq p} \delta_{ti} =$

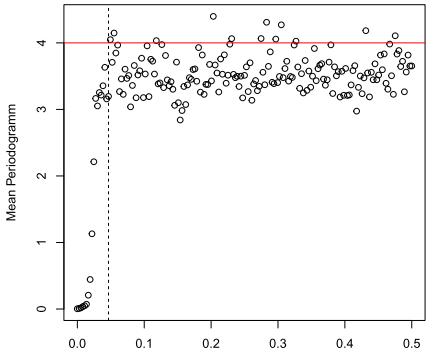

FIG 3. *Components of the averaged periodograms* $\hat{\xi} = T^{-1}\sum_{t=1}^{T} I_{\hat{U}_t}(\boldsymbol{\theta})$. *We use the simulation setting of Section 4.1 with* $U_{ti} \overset{iid}{\sim} N(0,4)$, $p = 365$ *and* $T = 200$. *The dotted line at* $c = 0.1$ *indicates the* 10% *of lowest fundamental frequencies.*

$O_P(p^{-1/2})$, does not suffice to guarantee the same asymptotic distribution as in the iid case. For example, if $\delta_{ti} = \frac{1}{\sqrt{p}}\cos(2\pi ki/p)$ then this noise is within the above margins and a simple calculation shows that $I_{\delta_t}(\theta_k) = 1/4$, while it is zero for the other fundamental frequencies. Hence, in such a case the contribution of the error δ_{ti} is non-negligible. In our real data experiments we overcome the problem by excluding $\ell \in \mathcal{F}$ if $\ell < c \times q$, where, e.g., $c = 0.1$.

3.2. A variant of the scree plot

Determining the number of factors is a difficult problem. As previously mentioned in Section 2.1, among the existing approaches the methods by Onatski [38] and Owen and Wang [39] were the most accurate in our context. In this section we would like to propose an empirical approach, which is a visual tool similar to the widely used scree plot from Cattell [9]. We recall that the classical scree plot is based on the eigenvalues $\hat{\gamma}_1, \hat{\gamma}_2, \ldots$ of the empirical covariance matrix $\frac{1}{T}YY'$. It shows the eigenvalues in descending order. A kink (or an 'elbow') in the graph, where the rate of descent drops, indicates the number of factors.

Instead of eigenvalues we propose to use the values of our test-statistics $\hat{\Lambda}_{\text{inf}}$ or $\hat{\Lambda}_{\text{fin}}$ established in Section 3.1. The logic behind is as follows: assume that the errors $(U_{ti}\colon 1 \le i \le p)$ are iid and suppose we fit a factor model with $\ell < L$ factors. Then, a certain amount of cross-sectional dependence still prevails in the residuals $(\hat{U}_{ti}\colon 1 \le i \le p)$, since the estimator does not yet fully account for the common component. Hence, underestimating L is likely to result in a large value for the test statistics. When increasing the number ℓ of factors included in the model, the cross-sectional dependence is expected to diminish and finally to drop to a baseline level, when ℓ surpasses the true L, i.e., when in principle we move from a dependent to an independent sequence. Since our estimators $\hat{X}_t(\boldsymbol{s})$ are robust to overestimation of L [see, e.g., 16], we expect the test statistics

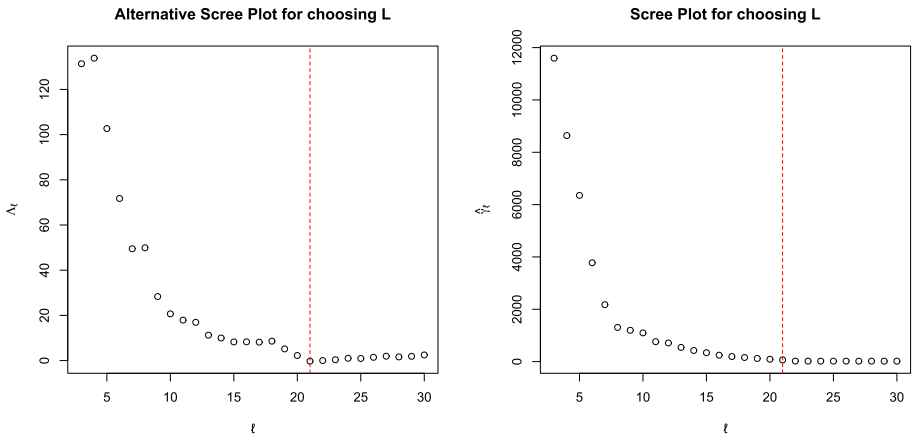

FIG 4. *Scree plots for choosing L in a simulation setting of Section 4.1. We use $U_{ti} \overset{iid}{\sim} N(0,4)$, $p = 365$ and $T = 200$. The true number of factors $L = 21$ is indicated by the dotted red line.*

to approximately remain constant for $\ell \geq L$. The method can be theoretically justified if the noise variables are iid, e.g., when we know that the noise can be related to measurement errors. In practice we may use it in a more general context. There we move from a long-range type dependence to weak dependence, which is likely to be reflected by a corresponding change in the decay rate of the test values.

We illustrate this approach in Figure 4, where we show plots of $\hat{\Lambda}_{\text{inf}}$ (figure on the left) and $\hat{\gamma}_\ell$ (figure on the right) against the chosen number of factors ℓ. Details of the related data is again provided in Section 4.1. We have chosen $T = 200$ and $p = 365$. These numbers are comparable to our real data example in Section 5. Due to the very large dimension, we are thinning out frequencies with $m = 20$ and as suggested in Section 3.1 and we also drop 10% of the lowest frequencies, so that $f/T \approx 0.04$. Given these parameters, it seems natural to employ $\hat{\Lambda}_{\text{inf}}$ (instead of $\hat{\Lambda}_{\text{fin}}$). In this example the true number of factors is $L = 21$ (marked by the dashed vertical line). This is also the value where our variant of the scree plot begins to approximately stay constant. From the standard scree plot we would deduce $\hat{L} = 8$ in this case.

In Figure 5 below we consider another setting, where $p = 48$ is relatively small compared to $T = 500$. In this example we simply use $\mathcal{F} = \{1, \ldots, q\}$. Since we get huge values for small ℓ we plot $\log \hat{\Lambda}^2_{\text{inf}}$. We can see that the 'scree' in our approach is much steeper and levels off near the true value of L. For the standard eigenvalues-based scree plot no accentuated kink can be spotted at $L = 21$. According to the 'elbow-rule', we would again chose $\hat{L} = 8$.

Remark 6. *While the eigenvalue based scree plot is monotone, this property cannot be guaranteed for our proposed scree plot.*

Remark 7. *The proposed method is based on the assumption of independent*

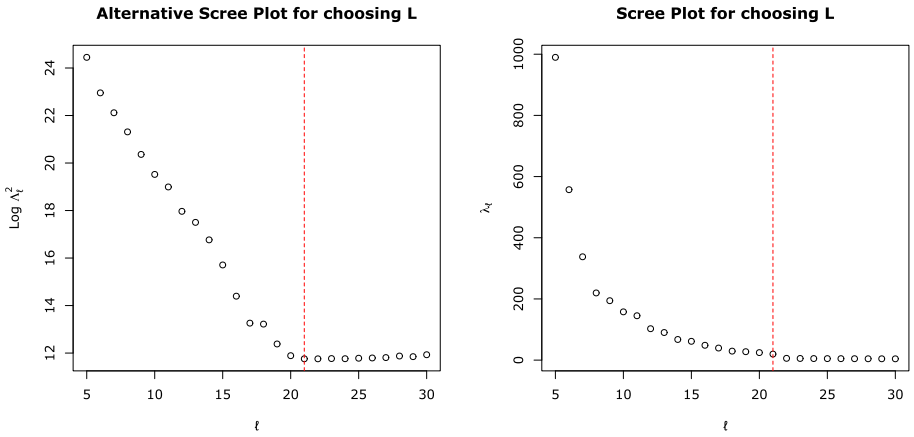

FIG 5. *Scree plots for choosing L in a simulation setting of Section 4.1. We use $U_{ti} \overset{iid}{\sim} N(0,4)$, $p = 48$ and $T = 500$. The true number of factors $L = 21$ is indicated by the dotted red line.*

noise. An important message of our paper is that in several real data examples the errors are not necessarily related to measurement errors and a certain degree of dependence is well expected. This is also the case for the data we consider in Section 5. A modification of the approach which allows for weakly dependent errors would be interesting, but is out of the scope of this paper and will be subject of future research.

4. Simulation experiments

In this section we investigate the performance of our methods on simulated data examples. We have performed extensive simulation studies that can be separated into two types: smooth data (Section 4.1) and data where the underlying signal and its derivative contains discontinuities (Section 4.2). The following simulations were performed in `R version 4.0.3` [41].

4.1. Recovering smooth signals

We consider bi-hourly measurements of particulate matter `pm10` in Graz from October 1st 2010 to March 31st 2011. Thus, we have 48 observations per day over the course of 182 days. To have control over the actual structure of the data, we generated synthetic curves by the following four steps: (1) transform the raw data to functional data; (2) create a bootstrap sample of size T thereof; (3) evaluate the resulting sample on a grid of intraday time points; (4) add noise as in (1.1). In Step (1) we chose to do a least squares fit using 21 cubic B-splines. This gives rise to relatively smooth curves. Then $T = 50, 100, 200, 500$ curves were obtained by the bootstrapping in Step (2). These curves are considered as

our signals $X_1(s), \dots, X_T(s)$. The signals in turn were evaluated at $p = 24, 48, 96$ equidistant points in $[0, 1]$, giving rise to $X_t(\boldsymbol{s})$ (Step (3)). In the final step we generated $Y_t = X_t(\boldsymbol{s}) + U_t$ with $U_t \overset{\text{iid}}{\sim} N_p(0, \Omega)$. For Ω we chose the covariance of a sample $(\varepsilon_1, \dots, \varepsilon_p)'$ from the stationary AR(1) process $\varepsilon_k = \theta\varepsilon_{k-1} + \xi_k$, where (ξ_k) is white noise with zero mean and variance σ^2. Hence $\Omega = \Omega(\theta, \sigma^2)$. We then consider $\Omega(0, 4)$, $\Omega(0, 16)$, $\Omega(0.4, 1)$ and $\Omega(0.8, 1)$.

Our goal is now to recover the signal $X_t(\boldsymbol{s})$. First we compare our proposed method with a B-spline (B) and penalized B-spline smoothing approach (B_{pen}). Since the actual signal in this simulation setting is already contained in a space spanned by B-splines, we consider in fact a setup which is favourable for these competitors. The B-spline smooth was computed using $p/3$ basis functions and methods from the `fda` package [43] in `R`. When $p = 48$ this yields a number which is comparable to the actual number of B-splines used to create the signal, otherwise it is bigger. This is in line with Wood [51], who suggests using more basis functions than one believes necessary and then using a penalization approach to smooth the result. We use penalized B-splines with a roughness penalty of the form $\int X''(s)^2 ds$. The penalty is added to the regular least squares equation and weighted with a parameter λ, which needs to be chosen. This has been done using a GCV (generalized cross validation) technique as described in Ramsay et al. [42].

Furthermore, we compare our approach to the functional principal components (FPC) approach as motivated in Staniswalis and Lee [46]. To this end, we have used the function `fpca.sc` from the `refund` package [24] in `R`, which smooths the empirical covariance prior to obtaining an estimate for the functional scores and subsequently, the estimated signal. The number of principal components was automatically chosen to be large enough to explain 99% of the variance. Note that in this approach, the smoothing of the covariance operator is done via penalized splines, which is in line with a suggestion in Di et al. [15]. The number of basis functions we used in this smoothing is $p/3$ as well. In our exploration we found that increasing the number of splines in this function requires immense computational effort while giving little improvement.

For the factor analysis, we used two different approaches. First, we used the PCA driven approach, as described in Fan et al. [16] and explained in our Section 2.1 (FA_{PCA}). Second, we use a Maximum-Likelihood approach (FA_{ML}) with the EM algorithm as described in Bai and Li [2] and implemented in the package `cate`. As for choosing the number of factors, we used the methods BCV and ED, which are described in Section 2.1. We note that the method we proposed in Section 3.2 provides a powerful visual tool, but choosing L in this way for hundreds of simulation runs is not practically feasible.

Implementation of BCV and ED can also be found in the package `cate` (see [50]). Note that for the implementation, a maximum number of factors `rmax` to be considered can be selected. We have chosen `rmax` $= 23$. Estimates tend to be robust to the overestimation of the dimension L, but sensitive to too small L, see for example Fan et al. [16]. This is intuitive, as a too small choice of L will result in important information being excluded from the fit, whereas we only add

potentially "insignificant" information if L is chosen too large. Thus, we have used $\hat{L} = \max(\hat{L}_{\mathrm{BCV}}, \hat{L}_{\mathrm{ED}})$. Practically we experience that in most settings $\hat{L}_{\mathrm{BCV}} \geq \hat{L}_{\mathrm{ED}}$. Hence the results remain basically unchanged if $\hat{L} = \hat{L}_{\mathrm{BCV}}$ is used.

In order to evaluate the quality of the respective approaches we are interested in the error $X_t(\boldsymbol{s}) - \hat{X}_t(\boldsymbol{s})$. While for real data $X_t(\boldsymbol{s})$ is not observable, the signals are known in our simulation setting and we can hence define

$$\mathrm{SSE}^{\mathrm{appr}} = \frac{1}{pT} \sum_{i=1}^{p} \sum_{t=1}^{T} (X_t(s_i) - \hat{X}_t(s_i))^2. \tag{4.1}$$

The results of our Monte Carlo study with 250 iterations can be found in Tables 1 and 2. Methods that produce the minimal $\mathrm{SSE}^{\mathrm{appr}}$ in each instance are bold.

Dimensions		$\mathrm{SSE}^{\mathrm{appr}}$ ($\sigma^2 = 4$)						$\mathrm{SSE}^{\mathrm{appr}}$ ($\sigma^2 = 16$)					
p	T	$\hat{L}$	B	$\mathrm{B}_{\mathrm{pen}}$	FPC	$\mathrm{FA}_{\mathrm{ML}}$	$\mathrm{FA}_{\mathrm{PCA}}$	$\hat{L}$	B	$\mathrm{B}_{\mathrm{pen}}$	FPC	$\mathrm{FA}_{\mathrm{ML}}$	$\mathrm{FA}_{\mathrm{PCA}}$
24	50	8	39.50	41.76	39.45	22.07	**16.59**	7	43.07	45.62	42.58	31.64	**27.07**
24	100	11	39.55	42.11	39.94	15.02	**10.33**	9	42.61	45.85	42.14	25.04	**21.18**
24	200	14	39.79	42.39	39.82	11.16	**7.32**	11	43.54	47	42.85	22.63	**18.4**
24	500	18	39.07	41.51	39.95	6.07	**4.34**	13	43.30	47.44	42.91	19.28	**14.93**
48	50	12	6.87	6.92	14.32	7.09	**5.51**	10	10.68	**10.66**	16.95	15.34	14.12
48	100	21	6.74	6.75	14.40	3.02	**2.62**	13	10.83	10.9	17.21	11.70	**10.82**
48	200	21	6.76	6.77	13.91	2.14	**2.1**	16	10.77	10.91	16.85	9.82	**9.08**
48	500	22	6.77	6.79	14.06	1.99	**1.96**	21	10.80	11.01	16.82	8.05	**7.87**
96	50	16	1.33	**1.19**	6.91	2.95	2.69	12	5.24	**4.13**	9.14	9.88	9.67
96	100	21	1.32	**1.18**	7.02	1.72	1.68	16	5.24	**4.14**	9.03	7.08	6.96
96	200	22	1.32	**1.18**	7.21	1.32	1.33	19	5.24	**4.17**	9.00	5.53	5.42
96	500	22	1.32	1.18	7.16	1.10	**1.09**	21	5.24	**4.2**	8.78	4.30	4.34

Table 1: Simulation results ($\mathrm{SSE}^{\mathrm{appr}}$) for the synthetic `pm10` data under settings $\Omega(0, 4)$ and $\Omega(0, 16)$. Here $\hat{L}$ is median value of the estimates $\max(\hat{L}_{\mathrm{BCV}}, \hat{L}_{\mathrm{ED}})$.

The most important observations are summarised below:

1. The factor model approach outperforms the B-splines largely when p is growing slower than T. The penalized B-splines work best if p is very large and T is small. In this case the noise can be very well smoothed on a local level.
2. As expected, for the B-splines based approaches the $\mathrm{SSE}^{\mathrm{appr}}$ does not decrease with growing sample size only with increasing p. Against our expectations, the FPC method did not improve in practice with increasing T either, though theoretically it should (see the results in Müller et al. [37]). It seems that the eigenfunctions from the smoothed covariances are oversmoothing the data and then local features of the data cannot be accurately recovered. In contrast, for both factor model estimators $\mathrm{SSE}^{\mathrm{appr}}$ decreases significantly with increasing T as well as increasing p.

Dimensions		SSE$^{\text{appr}}$ ($\theta = 0.4$)						SSE$^{\text{appr}}$ ($\theta = 0.8$)					
p	T	$\hat{L}$	B	B_{pen}	FPC	FA_{ML}	FA_{PCA}	$\hat{L}$	B	B_{pen}	FPC	FA_{ML}	FA_{PCA}
24	50	8	38.50	40.46	38.99	20.7	**14.61**	8	39.51	41.56	39.95	21.07	**15.15**
24	100	12	38.63	40.71	39.22	12.37	**7.34**	12	39.71	41.78	40.07	13.44	**8.51**
24	200	18	38.42	40.49	39.14	4.52	**2.88**	18	40.35	42.17	40.63	5.76	**4.19**
24	500	19	38.88	40.69	39.62	2.23	**1.63**	19	40.09	41.66	40.51	4.31	**3.47**
48	50	21	6.05	6.12	13.55	1.21	**1.08**	21	8.00	8.06	14.83	2.71	**2.66**
48	100	21	6.18	6.25	14.00	0.92	**0.91**	21	7.91	7.98	15.71	2.6	**2.58**
48	200	22	6.13	6.21	13.65	0.88	**0.87**	22	7.90	7.97	15.41	2.58	**2.56**
48	500	22	6.16	6.23	13.71	0.86	**0.85**	22	7.87	7.94	15.08	2.57	**2.56**
96	50	21	0.70	**0.69**	6.55	0.89	0.84	21	2.43	**2.42**	7.73	2.58	2.5
96	100	22	0.70	**0.69**	6.58	0.71	0.69	22	2.42	2.42	7.94	2.4	**2.36**
96	200	22	0.70	0.69	6.79	0.64	**0.62**	23	2.42	2.41	7.93	2.37	**2.32**
96	500	22	0.70	0.69	6.65	**0.56**	**0.56**	23	2.43	2.42	8.03	2.35	**2.29**

Table 2: Simulation results (SSE$^{\text{appr}}$) for the synthetic `pm10` data with AR(1) noise.

3. The FA_{PCA} approach gave better results than the FA_{ML} approach.

We have also experimented with further simulations settings. Not surprisingly, by further increasing σ^2, SSE$^{\text{appr}}$ increases for all methods. Nevertheless we observe that in comparison to each other the methods behave similarly as in the settings described. The combination large σ^2, large p and very small T (e.g., $T = 10$) favours our competitors, while our proposed approach improves considerably with growing T in all instances. For only mildly larger T and much larger p (e.g., $p = 96, 192$ and $T = 30$) we immediately obtain estimates that are competitive with the other approaches.

Since the signals in our simulations are relatively smooth, it is no surprise that smoothing methods perform well for large p. For curves with rough signal, smoothing approaches are not able to recover specific features of the signal due to oversmoothing. This is discussed in the following section.

4.2. Recovering signals with discontinuities

We check in the following simulation setting the practical impact of "rough" signals on the respective methods. More specifically, the signals $X_t(s)$ are defined on $[0, 1]$ and are constructed as follows:

$$X_t(s) = \sum_{k=1}^{3} \xi_{tk} \varphi_k(s),$$

where $\varphi_1(s) = \mathbb{1}_{\{s>1/3\}}$, $\varphi_2(s) = (-1)^{\kappa(s)} 4(0.2 - |s - 0.5|)\mathbb{1}_{\{s\in[1/3,2/3]\}}$ and $\varphi_3(s) = \cos 6\pi s$, where $\kappa(s) = \mathbb{1}_{\{s\in(1/2,2/3]\}}$. The associated scores are independent and normally distributed $\xi_{tk} \sim N(0, 2^{-2(k-1)})$ for $k = 1, 2, 3$. The noisy observations are obtained via $Y_{ti} = X_t(s_i) + U_{ti}$, where for U_{ti} we consider again

Gaussian processes as in the previous section. In particular we use covariances $\Omega(0, 0.01)$, $\Omega(0, 0.05)$ and $\Omega(0, 0.1)$. We consider equidistant observation points $s_i = (i - 0.5)/p$ for $i = 1, \ldots, p$. Thus, the signals may be disrupted at $s = 1/3$ (through φ_1), and they have a discontinuous derivative at $s = 1/2$ (through φ_2). Figure 6 shows two sample curves (black line) and the corresponding noisy observations (circles).

We consider the configurations $p = 20, 50, 70$ and $T = 50, 100, 200, 400$. With the three different σ^2's this gives rise to a total of 36 different settings, which have been repeated 200 times each. The signal is estimated by the methods $\mathrm{FA}_{\mathrm{PCA}}$, $\mathrm{B}_{\mathrm{pen}}$ and FPC. The rest of the procedure is the same as in Section 4.1. The results are summarized in Table 3.

Dimensions		$\mathrm{SSE}^{\mathrm{appr}}$ ($\sigma^2 = 0.01$)				$\mathrm{SSE}^{\mathrm{appr}}$ ($\sigma^2 = 0.05$)				$\mathrm{SSE}^{\mathrm{appr}}$ ($\sigma^2 = 0.1$)			
p	T	$\hat{L}$	$\mathrm{B}_{\mathrm{pen}}$	FPC	$\mathrm{FA}_{\mathrm{PCA}}$	$\hat{L}$	$\mathrm{B}_{\mathrm{pen}}$	FPC	$\mathrm{FA}_{\mathrm{PCA}}$	$\hat{L}$	$\mathrm{B}_{\mathrm{pen}}$	FPC	$\mathrm{FA}_{\mathrm{PCA}}$
20	50	4	5.6	4.6	**0.4**	3	7.2	5.0	**1.2**	3	9.4	5.6	**2.5**
20	100	5	5.7	4.6	**0.4**	3	7.2	5.1	**1.0**	3	9.7	5.5	**2.0**
20	200	5	5.7	4.6	**0.3**	3	6.9	4.9	**1.0**	3	9.7	5.5	**1.8**
20	400	5	5.7	4.5	**0.3**	3	6.9	4.9	**1.0**	3	9.8	5.4	**1.7**
50	50	5	1.5	1.1	**0.3**	3	2.6	1.4	**0.8**	3	3.5	1.8	**1.5**
50	100	5	1.5	1.1	**0.2**	3	2.6	1.3	**0.6**	3	3.4	1.6	**1.1**
50	200	3	1.5	1.0	**0.1**	3	2.6	1.3	**0.4**	3	3.4	1.6	**0.8**
50	400	3	1.5	1.0	**0.1**	3	2.6	1.3	**0.4**	3	3.4	1.6	**0.7**
70	50	5	1.1	0.6	**0.2**	3	2.0	0.9	**0.7**	3	2.8	**1.2**	1.4
70	100	3	1.1	0.6	**0.1**	3	2.0	0.8	**0.4**	3	2.8	1.0	**0.9**
70	200	3	1.1	0.6	**0.1**	3	2.0	0.8	**0.3**	3	2.8	1.0	**0.6**
70	400	3	1.1	0.6	**0.1**	3	2.0	0.7	**0.3**	3	2.8	1.0	**0.5**

Table 3: Simulation results ($100 \times \mathrm{SSE}^{\mathrm{appr}}$) for discontinuous signals under setups $\Omega(0, 0.01)$, $\Omega(0, 0.05)$ and $\Omega(0, 0.1)$.

$\mathrm{FA}_{\mathrm{PCA}}$ broadly outperforms its competitors. It is evident that the penalized B-spline as well as the FPC approach both fail to accurately estimate the signal at the discontinuity $s = 1/3$; see Figures 6 and 7.

We also mention that $\hat{L}$ can be overestimated as can be seen in the case of $\sigma^2 = 0.01$. Despite the mild overestimation of the required number of factors, we see no negative impact on the recovery of the signal.

We note that the function $\psi_1(s) = \sqrt{3/2}\varphi_1(s)$ is an eigenfunction of this process. As outlined in Section 2.3 we may estimate this eigenfunction from the raw data. Our estimate is subsequently compared to the first functional principal component obtained using the method motivated by Staniswalis and Lee [46] and implemented in the package `refund`. Furthermore, we compare our result to the principal components obtained from using the penalized B-spline model, using the package `fda`. The resulting estimates are shown in Figure 7. Both, FPC and $\mathrm{B}_{\mathrm{pen}}$ cannot appropriately recover the jump around $s = 1/3$. Our suggested approach recovers this particular feature very accurately.

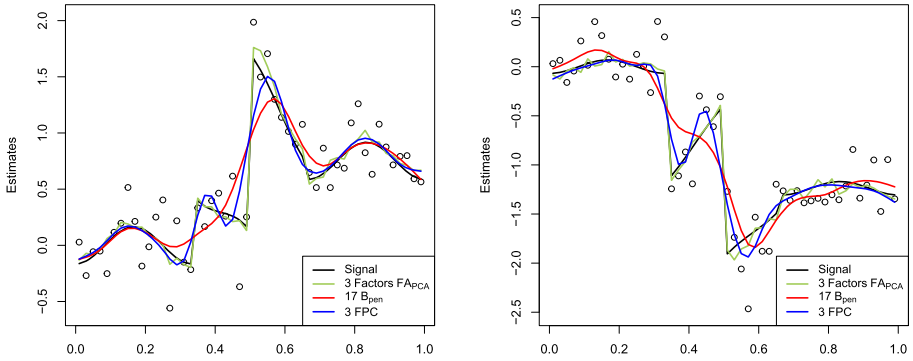

FIG 6. *Estimates for the rough signal simulation for* $p = 50, T = 200, \sigma^2 = 0.05$. *Dots represent the noisy observations.*

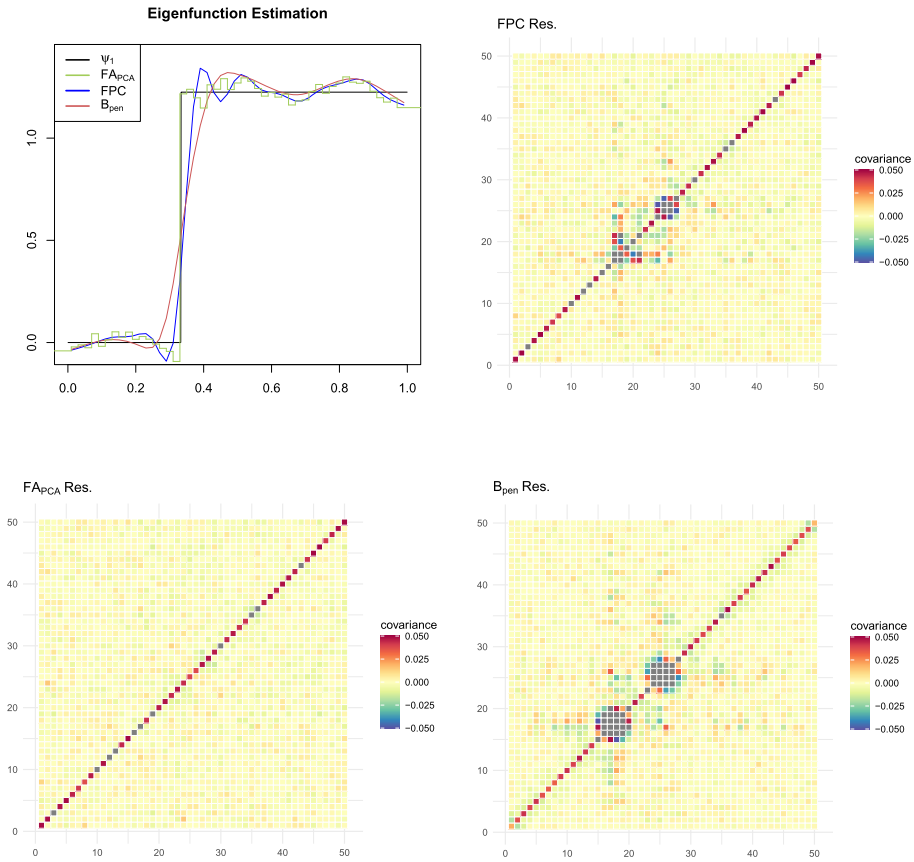

FIG 7. *Estimates for the first eigenfunction* $\psi_1(s)$ *(top left) and heat maps of the empirical covariance matrix of the residuals for the FPC (top right),* FA_{PCA} *(bottom left) and* B_{pen} *approach (bottom right) for the rough signal simulation.*

4.3. Testing for independent noise

In this section, we investigate the size and power of the tests $\hat{\Lambda}_{\text{inf}}$ and $\hat{\Lambda}_{\text{fin}}$ developed in Section 3.1. To this end, we consider the setting of Section 4.1, with $p = 365$ observations per curve for $T = 200$ curves. This sample size compares to the real data settings we consider in the next section. As for the errors we investigate $\Omega(0, 16)$ for determination of the size and $\Omega(0.05, 1)$ and $\Omega(0.1, 1)$ for the power. We then fit a factor model using the true number of factors $L = 21$ and test whether the model residuals are iid. As mentioned in Section 3.1, the spectral densities of the model residuals are not well-estimated at very low frequencies. This may be mitigated by only considering frequencies θ_ℓ with $\ell > cq$ for some small $c > 0$. Furthermore, our theoretical results only support the case of $f/T \longrightarrow 0$, where f is the number of frequencies considered. In order to justify this setting, we only include every m–th frequency in our test statistic. As a point of comparison we have also applied the testing procedure to the actual (latent) errors. Following Gasser et al. [22] we estimate the variance σ^2 in each instance by $\hat{\sigma}^2 := T^{-1}\sum_{t=1}[6(p-2)]^{-1}\sum_{j=2}^{p-1}[U_{t,j+1} + U_{t,j-1} - 2U_{t,j}]^2$. Each setting has been repeated 1000 times and we check how often the test rejects $\mathcal{H}_0$ at significance levels $0.01, 0.05, 0.1$. The results are displayed in Table 4. With iid variables the size matches the level very well. For the actual residual errors, the tests are slightly too sensitive, but give decent results if we use a not too dense set of frequencies, in particular avoiding frequencies around 0.

			$\hat{\Lambda}_{\text{inf}}$			$\hat{\Lambda}_{\text{fin}}$		
	level α		0.01	0.05	0.1	0.01	0.05	0.1
	c	m						
	0.10	10	0.046	0.098	0.145	0.022	0.090	0.146
	0.10	20	0.046	0.092	0.131	0.022	0.084	0.145
	0.20	10	0.035	0.079	0.137	0.018	0.065	0.123
	0.20	20	0.031	0.072	0.108	0.012	0.057	0.114
U_{Norm}	0.10	20	0.019	0.053	0.092	0.007	0.045	0.103

Table 4: Empirical test sizes using $\hat{\Lambda}_{\text{inf}}$ and $\hat{\Lambda}_{\text{fin}}$ on the setting $\Omega(0, 16)$.

Finally, we investigate the power of the test under the covariance settings $\Omega(0.1, 1)$ and $\Omega(0.4, 1)$. The results are shown in Table 5. We have only considered tests with $m = 20$ and $c = 0.2$, since here the size was closest to the nominal level. We see the very good power of our tests confirmed.

			$\hat{\Lambda}_{\text{inf}}$			$\hat{\Lambda}_{\text{fin}}$		
	level α		0.01	0.05	0.1	0.01	0.05	0.1
θ	c	m						
0.05	0.2	20	0.302	0.415	0.492	0.171	0.333	0.432
0.1	0.2	20	0.951	0.971	0.979	0.902	0.960	0.974

Table 5: Empirical power on the settings $\Omega(0.1, 1)$ and $\Omega(0.4, 1)$.

5. Real data illustrations

In the following two subsections we analyse annual temperature curves from Canada. We differentiate between two settings: On the one hand, we analyse a *temporal setting* in the sense that we consider curves over several years in one location. On the other hand, we consider a *spatial setting*, where we investigate the same year for different weather stations. Our objective is to transform daily mean temperature data throughout a year into annual temperature curves. The data was acquired from https://climate.weather.gc.ca/ and curves with more than 10% missing observations were discarded in both the temporal and spatial setting. Remaining missing observations were imputed using interpolation.

Ramsay et al. [42] have smoothed this type of data with 65 Fourier basis functions and a penalization term. We follow this route, but instead use a B-spline basis (also with 65 basis functions) and a roughness penalty of the form $\int (f''(x))^2 dx$ ($\mathrm{B}_{\mathrm{pen}}$). The tuning parameter controlling the size of the penalization term is chosen with generalized cross validation techniques as in Ramsay et al. [42]. The second method of comparison is FPC. These two approaches are then compared to $\mathrm{FA}_{\mathrm{PCA}}$.

5.1. Temporal Data: St. Margaret's Bay

We consider annual temperature curves from St. Margaret's Bay in Nova Scotia, Canada. This weather station has a long history of recorded data, from which we will use a selection of 91 yearly curves ranging from 1923 to 2020. The first goal is to determine the number of factors. We can see in Figure 8 that the regular scree plot, ED and BCV indicate $L = 3$. For our variant of the scree plot we use $\hat{\Lambda}_{\mathrm{inf}}$ with $c = 0.1$ and $m = 20$ and deduce from this $\hat{L} = 14$. The values of the test statistic remain very large for all choices of L, which indicates that the residuals are not iid.

The residual covariance with $L = 14$ is shown in Figure 9. In contrast to $\mathrm{B}_{\mathrm{pen}}$ (see Figure 2) we observe a strong concentration of the covariances on the diagonal and no oscillation in the empirical autocovariance function. Using FPC we got similar results as with $\mathrm{B}_{\mathrm{pen}}$.

For any of the approaches the residuals are obviously not iid and the tests discussed in Section 3.1 (using $c = 0.1$ and $m = 20$) clearly reject this hypothesis. The averaged periodogram ordinates $\frac{1}{T}\sum_{t=1}^{T} I_{\hat{U}}(\boldsymbol{\theta})$ which we compute for the residuals obtained from the three investigated methods can be used as estimators for the corresponding spectral densities (see Figure 10). Observe that we have a strong bias towards zero at frequencies close to 0 for all methods. This indicates that low frequencies are removed and attributed to the signals. This phenomenon is most pronounced for $\mathrm{B}_{\mathrm{pen}}$.

5.2. Spatial Data: Canadian weather stations

Now we consider a spatial setting, where each of the annual curves corresponds to a weather station. To this end, we have compiled the data from weather

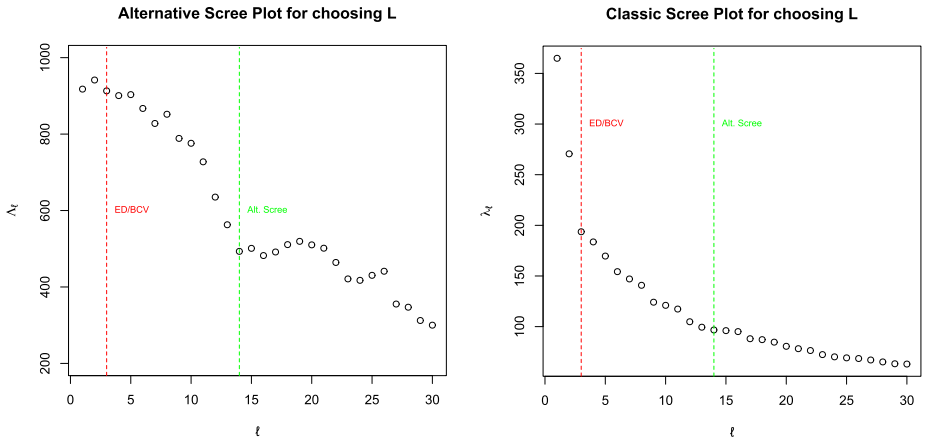

FIG 8. *Alternative (left) and classic (right) scree plot. The estimates* $\hat{L}_{ED}$ *and* $\hat{L}_{BCV}$ *are indicated by the red vertical lines.*

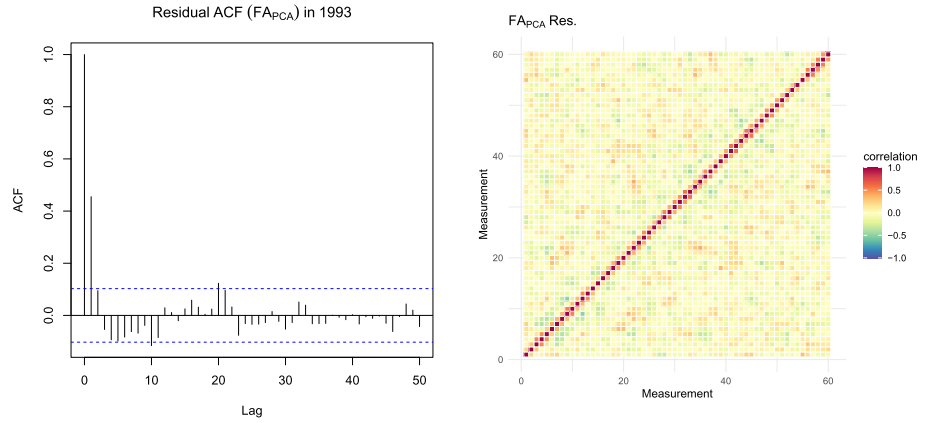

FIG 9. *We use* FA_{PCA} *and show the empirical autocorrelation function for the residuals in St. Margaret's Bay in 1993 (left) and a heat map of the empirical residual correlation matrices (right, restricted to the first two months of a year).*

stations in the provinces of Quebec and Ontario with daily mean temperature measurements available in 2013. After imputing scarcely scattered missing values and removing stations with too much missing data, we have $T = 213$ curves left. While the same general structure as in the temporal setup can be observed, we expect a different residual behavior. The idiosyncratic components now describe a station-specific error. Since spikes in the temperature curves are likely to occur across several stations, we expect a close co-movement resulting in much smaller idiosyncratic noise terms.

We begin by choosing the number of factors. The scree plot (see Figure 11)

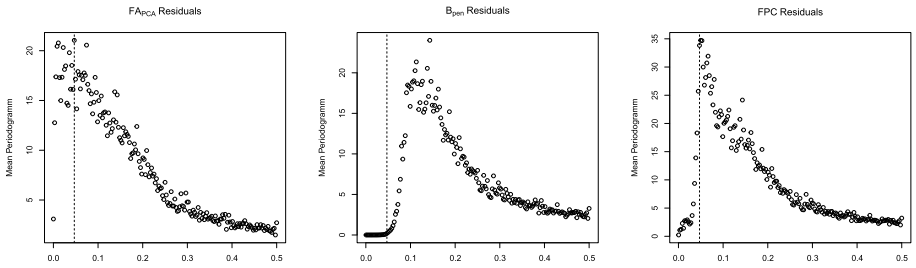

FIG 10. *Mean Periodogramm* $\xi = T^{-1}\sum_{t=1}^{T} I_{\hat{U}_t}(\boldsymbol{\theta})$ *for the* FA_{PCA} *(left),* B_{pen} *(middle) and FPC (right) residuals in St. Margaret's Bay Data. Vertical dotted lines indicate a proportion of* $c = 0.1$ *of low frequencies that we recommend to exclude for the statistics* $\hat{\Lambda}_{inf}$ *and* $\hat{\Lambda}_{fin}$.

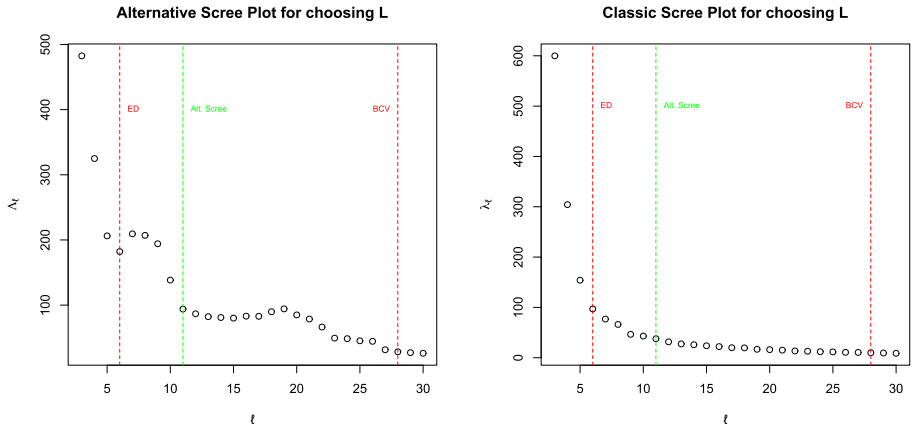

FIG 11. *Alternative (left) and classic (right) Scree plots to estimate the number of factors for the spatial data example. The estimates* $\hat{L}_{ED}$ *and* $\hat{L}_{BCV}$ *are indicated in red.*

indicates $\hat{L} = 6$, which coincides with the choice by the ED criterion, whereas BCV sets $\hat{L} = 27$. Our alternative scree plot (using again $c = 0.1$ and $m = 20$ for $\hat{\Lambda}_{\text{inf}}$) suggests $\hat{L} = 11$, and this is number we choose.

Looking at the residual covariances (Figure 12) we see that B_{pen} and FPC produce rather spurious results, whereas FA_{PCA} reasonably supports our assumptions on the idiosyncratic noise. Although the factor model has chosen fewer factors in the spatial setting compared to the temporal setting, we observe that now the signal follows the raw data more closely (see Figure 13) than we have conjectured above. In contrast, B_{pen} and FPC essentially produce the same results in the spatial and temporal setup. We indicate this in Table 6, where we show the estimated variance of the residuals.

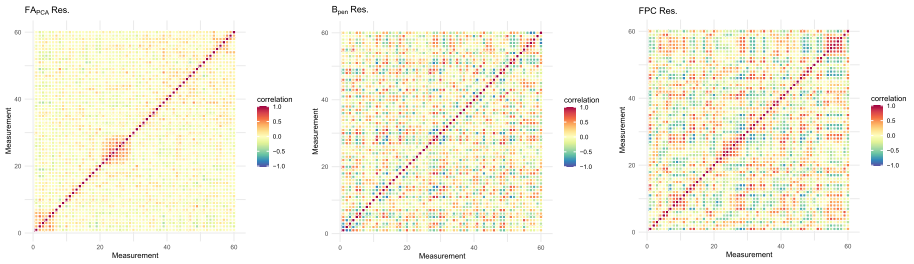

FIG 12. *Heat maps of the empirical correlation matrices for the FA_{PCA} (left), B_{pen} (middle) and FPC (right) approach for the spatial Canadian Weather Station Data. We show the first 60 days of the year.*

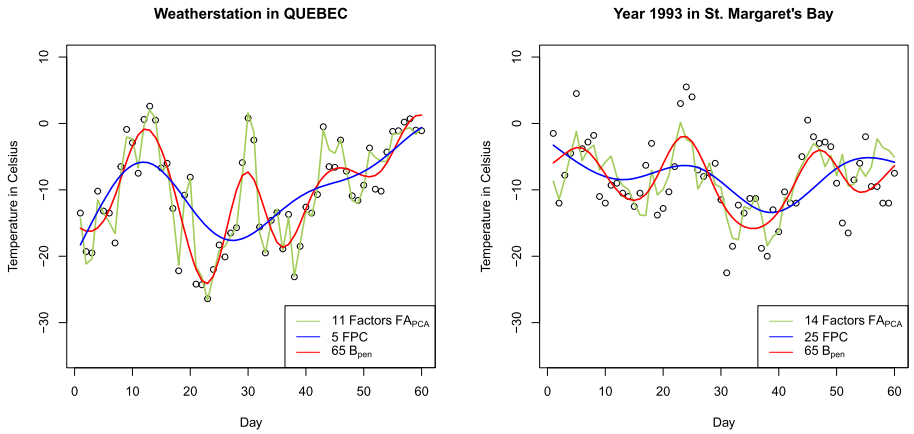

FIG 13. *Estimated signal for station Quebec from the spatial data (left) and estimated signal for station St. Margaret's Bay from the temporal data (right). Dots represent original observations. We show the first 60 days of the year.*

	FA_{PCA}	B_{pen}	FPC
Temporal	8.109	6.791	9.534
Spatial	1.718	5.046	7.351

Table 6: Estimated residual variances for the real life data examples.

6. Conclusion

In this paper we give a multivariate perspective to the modelling of discretely observed functional data. We outline that such data follow some approximate factor models which play an important role in macroeconomics. This perspective yields ready to use methods to estimate the latent signal without requiring smoothness of the curves. We show that this approach works extremely well on simulated data and leads to interesting results on real data. Moreover, this

paper offers some tools for analysing the model residuals. Typically those are assumed to be iid, but very often no residual analysis is done in order to justify this strong assumption. A theoretical foundation of the proposed estimation method is provided in our companion paper, Hörmann and Jammoul [29].

Appendix A: Appendix

A.1. Technical assumptions

Assumption 2. *The noise process (U_t) is iid zero mean and independent of the signals (X_t). The processes $(U_{ti}\colon 1 \le i \le p)$ are stationary and Gaussian with covariance function $\gamma^U(h) = Cov(U_{t,(i+h)}, U_{ti})$, such that $\sum_{h\in\mathbb{Z}} |\gamma^U(h)| \le C_U < \infty$.*

Assumption 3. *(a) The process $(X_t\colon t \ge 1)$ is zero mean and L^4-m-approximable. (b) The curves $X_t = (X_t(s)\colon s \in [0,1])$ define fourth order random processes (i.e., $\sup_{s\in[0,1]} EX_1^4(s) \le C_X < \infty$) with a continuous covariance kernel. (c) It holds that $E\sup_{s\in[0,1]} X_1^2(s) < \infty$. (d) Observations X_t lie in some L-dimensional function space.*

Assumption 4. *For the eigenfunctions φ_ℓ it holds that*

$$\max_{1\le k,\ell\le L} \left| p^{-1} \sum_{i=1}^{p} \varphi_k(s_i)\varphi_\ell(s_i) \right| = O(1)$$

as $T \to \infty$.

A.2. Proofs

We begin with an elementary lemma.

Lemma 1. *Let us denote by $\hat{X}_t(s)$, $s \in [0,1]$, the interpolation of the estimates $\hat{X}_t(s_i)$ as defined in (2.7) and let $\omega^f(\delta) = \sup_{s,s'\in[0,1]\colon |s-s'|\le\delta} |f(s) - f(s')|$ be the modulus of continuity of a function $f\colon [0,1] \to \mathbb{R}$. Then with $\delta = \max_{1\le i\le p-1} |s_{i+1} - s_i|$ we have*

$$\sup_{s\in[0,1]} |X_t(s) - \hat{X}_t(s)| \le 2\omega^{X_t}(\delta) + \max_{1\le i\le p} |X_t(s_i) - \hat{X}_t(s_i)|.$$

The lemma shows that the approximation error of the full curve can be decomposed into the modulus of continuity of the functional data and the approximation error on the observation grid. The proof of Lemma 1 can be easily seen and will thus be omitted.

Proof of Theorem 2.1. The main part of the proof essentially follows from the proof of Theorem 1 in Hörmann and Jammoul [29], noting that the bound

$$\max_{1\le i\le p} |X_t(s_i) - \hat{X}_t(s_i)| = O_P\left(\frac{1}{T^{1/4}} + \frac{T^{1/4}}{\sqrt{p}}\right)$$

obtained there can be easily improved. In fact, in this companion paper our bound for the estimation error involved an unnecessary uniform term $R^{(1)} = \max_t R_t^{(1)}$ instead of the specific $R_t^{(1)}$. (The variables $R_t^{(1)}$ are defined right after equation (8) in Hörmann and Jammoul [29].) Observing that the uniformity in t is not needed here, the proof reveals that

$$\max_{1\le i\le p} |X_t(s_i) - \hat{X}_t(s_i)| = O_P\left(R_t^{(1)} + R^{(2)}\right),$$

where it follows from Lemma 5 in Hörmann and Jammoul [29] that $R^{(2)} = O_P\left(T^{-1/2}\sqrt{\log p}\right)$. Following the arguments of Lemma 4 in Hörmann and Jammoul [29], where a bound for $R^{(1)}$ is derived, it can be readily shown that $R_t^{(1)} = O_P(T^{-1/2} + p^{-1/2})$.

For the modulus of continuity $\omega^{X_t}(\delta)$ we may conclude with Markov's inequality that

$$P(\omega^{X_t}(\delta) > \kappa\delta^\alpha) \le EM_t/\kappa.$$

Thus we see that $\omega^{X_t}(\delta) = O_P(\delta^\alpha)$ and the result follows using Lemma 1. □

Proof of Theorem 2.2. We decompose the $\|\varphi_\ell - \tilde{\varphi}_\ell\|$ into three pieces. To this end, we define the empirical covariance operator $\hat{\Gamma}^X$ of the fully observed $X_1, \ldots, X_T$ and its eigenfunctions $\hat{\varphi}_\ell$. Let $X_t^\star(s) := X_t(s_i)$ for $s \in [s_i, s_{i+1})$ be a discretized version of the fully observed data and let the associated empirical covariance operator be denoted by $\hat{\Gamma}^{X^\star}$ and its eigenfunctions by $\hat{\varphi}_\ell^\star$. Finally, let us define the empirical covariance matrix $\hat{\Sigma}^X = T^{-1}XX'$, where $X = (X_1(\boldsymbol{s}), \ldots, X_T(\boldsymbol{s}))$ and its associated eigenvectors $\hat{\psi}_\ell^X$. Consider

$$\|\varphi_\ell - \tilde{\varphi}_\ell\| \le \|\varphi_\ell - \hat{\varphi}_\ell\| + \|\hat{\varphi}_\ell - \hat{\varphi}_\ell^\star\| + \|\hat{\varphi}_\ell^\star - \tilde{\varphi}_\ell\|. \tag{A.1}$$

We may deduce from Weyl's theorem that

$$\|\varphi_\ell - \hat{\varphi}_\ell\| \le \frac{2\sqrt{2}}{\alpha_\ell}\|\Gamma^X - \hat{\Gamma}^X\|, \tag{A.2}$$

$$\|\hat{\varphi}_\ell - \hat{\varphi}_\ell^\star\| \le \frac{2\sqrt{2}}{\hat{\alpha}_\ell}\|\hat{\Gamma}^X - \hat{\Gamma}^{X^\star}\|, \tag{A.3}$$

where $\hat{\alpha}_\ell = \min\{\hat{\lambda}_\ell - \hat{\lambda}_{\ell+1}, \hat{\lambda}_{\ell-1} - \hat{\lambda}_\ell\}$ and where $\hat{\lambda}_\ell$ are the empirical eigenvalues of the fully observed data. From Hörmann and Kokoszka [30] it follows under Assumption 3 (a) that (A.2) is $O_P\left(T^{-1/2}\right)$ and that $\hat{\alpha}_\ell \to \alpha_\ell > 0$, as $T \to \infty$. Note that when $a(s,t)$ is the kernel of the bounded linear operator A, then $\|A\|^2 \le \int_0^1 \int_0^1 a^2(t,s)dsdt$. Hence

$$\|\hat{\Gamma}^X - \hat{\Gamma}^{X^\star}\|^2 \le \int_0^1 \int_0^1 \left(\frac{1}{T}\sum_{t=1}^T (X_t(r)X_t(s) - X_t^\star(r)X_t^\star(s))\right)^2 drds.$$

In the proof of Lemma 1 in Hörmann and Jammoul [29] it is shown that (2.10) implies that the right hand side is $O_P\left(p^{-1}\right)$ as $p \to \infty$. We hence conclude that (A.3) is $O_P\left(p^{-1/2}\right)$.

In the final step, we have to move from the functional setting to the matrix setting. It can be readily seen that the eigenvector $\hat{\psi}_\ell^X$ of $\hat{\Sigma}^X$ satisfies $\sqrt{p}[\hat{\psi}_\ell^X]_i = \hat{\varphi}_\ell^y(s)$ for $s \in [(i-1)/p, i/p)$. Thus, we may rewrite the last term in (A.1) as

$$\|\hat{\varphi}_\ell^y - \tilde{\varphi}_\ell\|^2 = (\hat{\psi}_\ell^X - \hat{\psi}_\ell^y)'(\hat{\psi}_\ell^X - \hat{\psi}_\ell^y).$$

Again by Weyl's theorem the right hand side is $O_P(\frac{1}{\hat{\beta}_\ell}\|\hat{\Sigma}^y - \hat{\Sigma}^X\|)$, where $\hat{\beta}_\ell = \min\{\hat{\gamma}_\ell - \hat{\gamma}_{\ell+1}, \hat{\gamma}_{\ell-1} - \hat{\gamma}_\ell\}$. Lemma 1 in Hörmann and Jammoul [29] implies that under Assumptions 2 and 3(a) and (b) we have

$$\hat{\gamma}_\ell - \hat{\gamma}_{\ell+1} \sim p(\lambda_\ell - \lambda_{\ell+1}).$$

Moreover, it is shown in this lemma that

$$\|\hat{\Sigma}^y - \hat{\Sigma}^X\| = \begin{cases} O_P(\sqrt{p}), & \text{if } p/T \to \gamma \in [0, \infty); \\ O_P(p/\sqrt{T}), & \text{if } p/T \to \infty. \end{cases}$$

Combining all bounds yields the desired convergence. □

Proof of Proposition 3.1. Suppose that the $Z = (Z_1, \ldots, Z_p)'$ has iid components with $EZ_1 = 0$ and $EZ_1^2 = \sigma^2$ and $EZ_1^4 < \infty$. Denote $\kappa := EZ_1^4 - 3$. Then it is well known that for any fundamental frequency we have that $EI_Z(\theta_\ell) = \sigma^2$ and

$$\mathrm{Cov}(I_Z(\theta_\ell), I_Z(\theta_{\ell'})) = \begin{cases} \sigma^4\kappa/p + \sigma^4 & \text{if } \ell = \ell'; \\ \sigma^4\kappa/p & \text{else.} \end{cases}$$

We thus have that the random vectors $V_t = I_{U_t}(\boldsymbol{\theta}) - \sigma^2 1_f$ are iid zero-mean and $\Sigma := \mathrm{Var}(V_t) = \sigma^4(I_f + \frac{\kappa}{p}\mathbb{1}_f) \in \mathbb{R}^{f\times f}$ holds. Consider the centering matrix $P_f := I_f - f^{-1}\mathbb{1}_f$ ($\mathbb{1}_f$ is the matrix with entries equal to 1) and note that

$$TS_\xi^2 = (f-1)^{-1}\left\|T^{-1/2}\sum_{t=1}^T P_f V_t\right\|^2.$$

We also note that $P_f\Sigma = \sigma^4 P_f$ and recall the well known fact that P_f has $f-1$ non-zero eigenvalues which are all equal to 1. If Q denotes the orthogonal matrix which has in its columns the related eigenvectors, then

$$TS_\xi^2 = (f-1)^{-1}\left\|T^{-1/2}\sum_{t=1}^T Q'P_f V_t\right\|^2 = (f-1)^{-1}\left\|T^{-1/2}\sum_{t=1}^T W_t\right\|^2,$$

where $(W_t', 0)' := Q'P_f V_t$. The vector W_t is zero-mean and $\mathrm{Var}(W_t) = \sigma^4 I_{f-1}$. By the central limit theorem the expression inside the norm converges to a

normally distributed vector with variance $\sigma^4 I_{f-1}$. The weak convergence of Λ_{inf} then follows by the continuous mapping theorem and Slutsky's lemma.

For growing f we consider the variable $\Lambda_{\text{inf}} = (TS_\xi^2/\sigma^4 - 1)\sqrt{(f-1)/2}$ and we wish to compare its distribution to the normal distribution, with its distribution function denoted by $\Phi(z)$. To this end let $Z \sim N(\mathbf{0}, I_{f-1})$ be a $(f-1)$-variate standard normal random vector. For any $z \in \mathbb{R}$ we get by the central limit theorem that

$$\left| P\left(\left(\frac{1}{(f-1)} \|Z\|^2 - 1 \right) \sqrt{\frac{(f-1)}{2}} \leq z \right) - \Phi(z) \right| \to 0.$$

Hence, it suffices to show that for all real z we have

$$\left| P(\Lambda_{\text{inf}} \leq z) - P\left(\left(\frac{1}{(f-1)} \|Z\|^2 - 1 \right) \sqrt{\frac{(f-1)}{2}} \leq z \right) \right| \to 0.$$

By Slutsky's lemma we can replace $\hat{\sigma}^4$ in the definition of Λ_{inf} by σ^4. With $\tilde{z} = \sqrt{(\frac{z\sqrt{2}}{\sqrt{f-1}} + 1)(f-1)}$ we hence need to show that

$$\left| P\left(\left\| T^{-1/2} \sum_{t=1}^{T} W_t/\sigma^2 \right\| \leq \tilde{z} \right) - P(\|Z\| \leq \tilde{z}) \right|. \tag{A.4}$$

If we can show that $E|W_{ti}|^4$ are uniformly bounded (in t and i), then by Corollary 3.1 in Fang and Koike [17] we get that for any $\tilde{z}$ the term (A.4) is bounded by

$$C\left(T^{-1/8} + (f/T)^{1/6} \right),$$

for some constant C which is independent of T and f. Thus we can guarantee convergence if $f/T \to 0$.

We want to show that $\max_{1 \leq i \leq f-1} \max_{1 \leq t \leq T} \mathrm{E}|W_{ti}|^4 < C$, where C does not depend on the dimension parameters f, p and T. It holds that $W_{ti} = v_i' V_t$, where v_i denotes the i-th column of the matrix Q and is thus an eigenvector of P_f belonging to a non-zero eigenvalue. It can be easily checked that for $f \geq 3$ the v_i can be written as $(0, \ldots, 1/\sqrt{2}, 0, \ldots, 0, -1/\sqrt{2})'$, with non-zero entries at the i-th and the last coordinate. Since we assume iid noise $(U_t,\ t \geq 1)$, it follows that $(W_{ti},\ t \geq 1)$ are iid as well and thus the expectations do not depend on t. Hence, let us consider $\mathrm{E}|v_i' V|^4$, where $V = (I_U(\theta_{j_1}), \ldots, I_U(\theta_{j_f}))' - \sigma^2 1_f$, with $\{j_1, \ldots, j_f\} = \mathcal{F}$ and $U = (u_1, \ldots, u_p)' \sim U_1$. We assume without loss of generality that $\sigma^2 = 1$. Then the k-th component of V is given by

$$V_k^c + V_k^s := \frac{1}{p} \left(\sum_{r=1}^{p} u_r \cos(\theta_{j_k} r) \right)^2 - 1/2 + \frac{1}{p} \left(\sum_{r=1}^{p} u_r \sin(\theta_{j_k} r) \right)^2 - 1/2.$$

We have

$$E(v'V^c + v'V^s)^4 \leq 16 \left(E(v'V^c)^4 + E(v'V^s)^4 \right)$$

$$\leq 64\left(E(V_i^c)^4 + E(V_f^c)^4 + E(V_i^s)^4 + E(V_f^s)^4\right).$$

All the terms on the right can be bounded in the same way. Let us consider $E(V_k^c)^4$. Noting that $\theta_{j_k} r = \theta_r j_k$ and $\sum_{r=1}^p \cos^2(\theta_r j_k) = p/2$ it can be written as

$$E(V_k^c)^4 = \frac{1}{p^4} E\left(\left(\sum_{r=1}^p u_r \cos(\theta_r j_k)\right)^2 - p/2\right)^4$$
$$\leq \frac{1}{p^4} E\left(\sum_{r=1}^p u_r \cos(\theta_r j_k)\right)^8.$$

The last inequality follows from the fact, that $E(X - EX)^4 \leq EX^4$ when X is a positive random variable. Now apply the Rosenthal inequality (see, e.g., [40]). □

Proof of Proposition 3.2. For the proof it suffices to show that $P(S_\xi^2 > \delta/2) \to 1$, $T \to \infty$. Now we have

$$S_\xi^2 = S_g^2 + S_{\xi-g}^2 + 2S_{g,\xi-g} \geq S_g^2 - 2|S_{g,\xi-g}|$$

where $S_{\xi-g}^2$ is defined analogously to S_ξ^2 and where

$$S_{g,\xi-g} = \frac{1}{f-1}\sum_{j=1}^f (g(\theta_{\ell_j}) - \bar{g})(\xi_j - g(\theta_{\ell_j}) - (\bar{\xi} - \bar{g}))$$
$$= \frac{1}{f-1}\sum_{j=1}^f (g(\theta_{\ell_j}) - \bar{g})(\xi_j - g(\theta_{\ell_j})).$$

By (3.3) $S_g^2 > \delta$ and it remains to show that $|S_{g,\xi-g}| \to 0$ in probability for $p \to \infty$. It is easy to see that $S_g^2 \leq 2\left(\sum_{h\in\mathbb{Z}} |\gamma_U(h)|\right)^2 < \infty$ and by Markov's inequality we have $|S_{g,\xi-g}|^2 \leq S_g^2 S_{\xi-g}^2$. Hence the claim follows if we can show that $S_{\xi-g}^2 \to 0$ in probability.

To this end we recall that

$$\sup_{\theta\in[-\pi,\pi]} |EI_{U_t}(\theta) - g(\theta)| \to 0 \quad (p \to \infty). \tag{A.5}$$

(See e.g. Proposition 10.3.1 in Brockwell and Davis [8].) Hence, when p is large enough, we have

$$P\left(\max_{\ell_j\in\mathcal{F}} |\xi_j - g(\theta_{\ell_j})| > \varepsilon\right) \leq \sum_{\ell\in\mathcal{F}} P\left(\left|\frac{1}{T}\sum_{t=1}^T (I_{U_t}(\theta_\ell) - g(\theta_\ell))\right| > \varepsilon\right)$$
$$\leq f \max_{\ell\in\mathcal{F}} P\left(\left|\frac{1}{T}\sum_{t=1}^T (I_{U_t}(\theta_\ell) - EI_{U_t}(\theta_\ell))\right| + \frac{1}{T}\sum_{t=1}^T |EI_{U_t}(\theta_\ell)) - g(\theta_\ell)| > \varepsilon\right)$$

$$\leq f \max_{\ell \in \mathcal{F}} P\left(\left|\frac{1}{T}\sum_{t=1}^{T}(I_{U_t}(\theta_\ell) - EI_{U_t}(\theta_\ell))\right| > \varepsilon/2\right)$$

$$\leq \frac{4f}{\varepsilon^2} \max_{\ell \in \mathcal{F}} E\left(\frac{1}{T}\sum_{t=1}^{T}(I_{U_t}(\theta_\ell) - EI_{U_t}(\theta_\ell))\right)^2$$

$$\leq \frac{4f}{\varepsilon^2 T^2} \max_{\ell \in \mathcal{F}} \sum_{t=1}^{T}\sum_{t'=1}^{T} \mathrm{Cov}(I_{U_t}(\theta_\ell), I_{U_{t'}}(\theta_\ell)) = \frac{4f}{\varepsilon^2 T} \max_{\ell \in \mathcal{F}} \mathrm{Var} I_{U_t}(\theta_\ell) \to 0. \qquad \square$$

Acknowledgments

We thank Jeff Goldsmith and Sonja Greven for a very helpful discussion on the `refund` package which we used to implement the FPC method. We also would like to thank two anonymous referees for a number of constructive remarks and a detailed list of corrections which helped to improve the paper.

References

[1] J. Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71:135–171, 2003. MR1956857

[2] J. Bai and K. Li. Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40:436–465, 2012. MR3014313

[3] J. Bai and K. Li. Maximum likelihood estimation and inference for approximate factor models of high dimension. *The Review of Economics and Statistics*, 98:298–309, 2016.

[4] J. Bai and Y. Liao. Efficient estimation of approximate factor models via regularized maximum likelihood. *Journal of Econometrics*, 191:1–18, 2016. MR3434432

[5] J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70:191–221, 2002. MR1926259

[6] M. Benko, W. Härdle, and A. Kneip. Common functional principal components. *The Annals of Statistics*, 37:1–34, 2009. MR2488343

[7] D. Bosq. *Linear processes in function spaces: theory and applications*. Lecture Notes in Statistics. Springer, New York, 2000. MR1783138

[8] P. Brockwell and R. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer New York, second edition, 1991. MR2839251

[9] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1:245–276, 1966.

[10] G. Chamberlain and M. Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51:1281–1304, 1983. MR0736050

[11] I. Choi. Efficient estimation of factor models. *Econometric Theory*, 28:274–308, 2012. MR2913632

[12] G. Claeskens, T. Krivobokova, and J. Opsomer. Asymptotic properties of penalized spline estimators. *Biometrika*, 96:529–544, 2009. MR2538755

[13] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979. MR0556476

[14] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12136–154, 1982. MR0650934

[15] C.-Z. Di, C. Crainiceanu, B. Caffo, and N. Punjabi. Multilevel functional principal components. *The Annals of Applied Statistics*, 3458–488, 2009. MR2668715

[16] J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B.*, 75603–680, 2013. MR3091653

[17] X. Fang and Y. Koike. Large-dimensional central limit theorem with fourth-moment error bounds on convex sets and balls. *https://arxiv.org/abs/2009.00339*, 2021.

[18] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice.* Springer-Verlag, Berlin, Heidelberg, 2006. MR2229687

[19] M. Forni and M. Lippi. The generalized dynamic factor model: representation theory. *Econometric Theory*, 17:1113–1141, 2001. MR1867540

[20] M. Forni, L. Reichlin, M. Hallin, and M. Lippi. The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics*, 82:540–554, 2000. MR1867540

[21] M. Forni, M. Hallin, Lippi M., and L. Reichlin. The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100:830–840, 2005. MR2201012

[22] T. Gasser, L. Sroka, and C. Jennen-Steinmetz. Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73:625–633, 12 1986. MR0897854

[23] I. Gohberg, S. Goldberg, and M.A. Kaashoek. *Basic Classes of Linear Operators*. Birkhäuser, 2000. MR2015498

[24] J. Goldsmith, F. Scheipl, L. Huang, J. Wrobel, C. Di, J. Gellar, J. Harezlak, M. W. McLean, B. Swihart, L. Xiao, C. Crainiceanu, and P. T. Reiss. *refund: Regression with Functional Data*, 2021. `https://CRAN.R-project.org/package=refund`. R package version 0.1-24.

[25] P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:109–126, 2006. MR2212577

[26] P. Hall and J. Opsomer. Theory for penalised spline regression. *Biometrika*, 92:105–118, 03 2005. MR2158613

[27] M. Hallin and R. Liška. Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102:603–617, 2007. MR2325115

[28] W. Härdle, P. Hall, and J.S. Marron. How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association*, 83:86–95, 1988. MR0941001

[29] S. Hörmann and F. Jammoul. Consistently recovering the signal from noisy functional data. *Journal of Multivariate Analysis*, 2022. 189, 2022. MR4384127

[30] S. Hörmann and P. Kokoszka. Weakly dependent functional data. *The Annals of Statistics*, 38:1845–1884, 2010. MR2662361

[31] L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Springer Science and Business Media, 2012. MR2920735

[32] O. Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002. MR1876169

[33] J. Kleffe. Principal components of random variables with values in a seperable hilbert space. *Mathematische Operationsforschung und Statistik*, 4:391–406, 1973. MR0391402

[34] P. Kokoszka and M. Reimherr. *Introduction to functional data analysis*. Chapman and Hall/CRC, 2017. MR3793167

[35] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press London; New York, 1979. MR0560319

[36] H.-G. Müller. Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, 82:231–238, 1987. MR0883351

[37] H.-G. Müller, U. Stadtmüller, and F. Yao. Functional variance processes. *Journal of the American Statistical Association*, 101:1007–1018, 2006. MR2324140

[38] A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92:1004–1016, 2010.

[39] A. Owen and J. Wang. Bi-cross-validation for factor analysis. *Statistical Science*, 31:119–139, 2016. MR3458596

[40] V.V. Petrov. *Limit Theorems of Probability Theory*. Oxford Science Publications, New York, 1995. MR1353441

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. https://www.R-project.org/.

[42] J. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and MATLAB*. Springer, 1st edition, 2009. MR3645102

[43] J. O. Ramsay, G. Hooker, and S. Graves. *fda: Functional Data Analysis*, 2021. https://CRAN.R-project.org/package=fda. R package version 5.5.1.

[44] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, New York, 2005. MR2168993

[45] T. Rubín and V. M. Panaretos. Sparsely observed functional time series: estimation and prediction. *Electronic Journal of Statistics*, 14:1137–1210, 2020. MR4069991

[46] J. Staniswalis and J. Lee. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93:1403–1418, 1998. MR1666636

[47] J. Stock and M. Watson. Macroeconomic forecasting using diffusion in-

dexes. *Journal of Business & Economic Statistics*, 20:147–162, 2002. MR1963257

[48] J. Stock and M. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97:1167–1179, 2002. MR1951271

[49] M.P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall, London, 1995. MR1319818

[50] J. Wang and Q. Zhao. *cate: High Dimensional Factor Analysis and Confounder Adjusted Testing and Estimation*, 2020. https://CRAN.R-project.org/package=cate. R package version 1.1.1.

[51] S. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017. MR2206355

[52] F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590, 2005. MR2160561