Econometric Theory, **31**, 2015, 1359–1381.
doi:10.1017/S0266466614000784

A NONPARAMETRIC ESTIMATOR FOR THE COVARIANCE FUNCTION OF FUNCTIONAL DATA

ALESSIO SANCETTA
Royal Holloway University of London

Many quantities of interest in economics and finance can be represented as partially observed functional data. Examples include structural business cycle estimation, implied volatility smile, the yield curve. Having embedded these quantities into continuous random curves, estimation of the covariance function is needed to extract factors, perform dimensionality reduction, and conduct inference on the factor scores. A series expansion for the covariance function is considered. Under summability restrictions on the absolute values of the coefficients in the series expansion, an estimation procedure that is resilient to overfitting is proposed. Under certain conditions, the rate of consistency for the resulting estimator achieves the minimax rate, allowing the observations to be weakly dependent. When the domain of the functional data is $K(>1)$ dimensional, the absolute summability restriction of the coefficients avoids the so called curse of dimensionality. As an application, a Box–Pierce statistic to test independence of partially observed functional data is derived. Simulation results and an empirical investigation of the efficiency of the Eurodollar futures contracts on the Chicago Mercantile Exchange are included.

1. INTRODUCTION

This paper is concerned with the estimation of the covariance function of functional data. The estimation framework is via series expansion. The contributions are as follows: (1) a constrained estimation method that allows to have good control of the estimation error; (2) a greedy algorithm to solve the estimation problem; (3) minimax rates of convergence for sparsely observable functional data when the true covariance function satisfies certain constraints, and avoidance of the curse of dimensionality when the domain of the covariance function is high dimensional; (4) using the estimated covariance function, a derivation of a Box–Pierce statistic for partially observed functional data. Regarding points (1)–(3), under the validity of certain semiparametric restrictions, the rates of convergence for sparsely observable functional data derived here are minimax and cannot be improved without additional conditions. Regarding point (4), this paper extends

I thank the anonymous referees and the co-editor Oliver Linton for comments that led to substantial improvements both in content and presentation. Address correspondence to Alessio Sancetta, Department of Economics, Royal Holloway University of London, Egham TW20 0EX, UK; e-mail: asancetta@gmail.com.

© Cambridge University Press 2014

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

a Box–Pierce kind of test for fully observable functional data (e.g., Horváth and Kokoszka, 2012, Thm. 2, Chap. 7) to the nonfully observable framework.

Functional data are realizations of functional random variables. In practice, we may observe discrete observations from partially observed random curves. For definiteness, consider $X_r(s_{rk})$ to be the r^{th} random function observed at s_{rk}, $k = 1, 2, \ldots, m_r$, where the observation points s_{rk}'s can be themselves random, in which case write S instead of s. In most applications, r represents the time dimension which is discrete, while S_{rk} a random point on the r^{th} curve, and in some cases one can think of k as indexing the cross-sectional dimension. This means that at each time r, instead of observing a vector of random variables, one could observe an entire smooth curve, or just some points on the curve and these points may change with r. The power of the framework rests on the fact that for each r, the data are like vectors that are highly correlated and this leads to improved asymptotics. To highlight the concept consider an i.i.d. (independent and identically distributed) sample $X_1(s), X_2(s), \ldots, X_n(s)$ of functional data observed at m points $s_1, s_2, \ldots, s_m \in [0,1]$, the same for all observation and deterministic for simplicity. Hence each observation is an m dimensional random vector, i.e., $[X_r(s_1), \ldots, X_r(s_m)]$ for the r^{th} observation. Given that the observations are functional data, by the law of large numbers in Banach spaces (e.g., Bosq, 2000),

$$\frac{1}{n}\sum_{r=1}^{n}(1-\mathbb{E})[X_r(s_1), \ldots, X_r(s_m)] \to [0, \ldots, 0]$$

a.s. (almost surely). If as $m \to \infty$ $s_1, s_2, \ldots, s_m$ becomes dense in $[0,1]$ the above remains true for functional data no matter how large m is, e.g., even if $m/n \to \infty$. Actually, the problem becomes easier from a statistical point of view when $m/n \to \infty$, as long as also $n \to \infty$. Contrast this with the usual law of large numbers for random vectors. By Chebyshev's inequality one knows that the above display is $O_p\left(\sqrt{m/n}\right)$, which does not go to zero unless $m = o(n)$. The smoothness implied by the functional data allows such powerful results. The covariance function $Cov(X_r(s), X_r(t))$ determines the smoothness of the functional data and it is a crucial ingredient for inference.

There has been some direct interest in the econometric literature, both theoretical and applied, on the use of functional data (e.g., Ramsay and Ramsey, 2002; Bugni, Hall, and Horowitz, 2009; Müller, Rituparna, and Stadtmüller, 2011; Bugni, 2012; Ferraty, Quintela-del-Río, and Vieu, 2012). Obvious examples of functional data in finance and economics are the yield curve, the implied volatility surface of derivative instruments, the volume weighted average price (VWAP), intradaily volatility processes, and many others.

Suppose that $P_r(s)$ is the price of a zero coupon bond on day r with maturity in s years. On each day r, we may only observe zero coupon prices for some maturities. These maturities change from day to day. Not every day new bonds with the same maturities of the ones outstanding the day before are usually issued. Then, at time r, $X_r(S_{rk}) = S_{rk}^{-1} \ln P_r(S_{rk})$ is the yield for the zero coupon

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

bond with maturity in S_{rk} periods, i.e., S_{rk} is the tenor (e.g., Brigo and Mercurio, 2006, p. 6).

Suppose that $X_r(s)$ is the implied volatility of a foreign exchange (Fx) European call so that s is the strike price. Call options for spot Fx are quoted in terms of implied volatility and the price is derived using some standard market convention and the Garman and Kohlhagen pricing formula. Usually, only the implied volatility for at the money options and the 25% and 75% delta options are widely quoted (e.g., Bisesti, Castagna, and Mercurio, 2005, for the terminology). This means that only 3 strike prices are very liquid, with the others being obtained by interpolation. The strikes change every day as the price of the Fx pair changes. Hence, S_{r1}, S_{r2}, and S_{r3} represent strike prices that change every day r.

The domain of the functional data is not always a subset of the real line as in the previous examples. It can be more complex. Suppose that $X_r(s)$ is a point in the swaption volatility surface. A swaption is an option on an interest rate swap (e.g., Brigo and Mercurio, 2006). Then, s is two dimensional. One dimension identifies the tenor (the length of the interest rate swap), while the other identifies the maturity of the swaption (when the swap starts).

The volume weighted average price (VWAP) is a critical quantity for optimal order execution (e.g., Almgren and Chriss, 2000; Almgren, 2003; Hasbrouk, 2007, for concepts and details). Let $P_r(S_{rk})$ be the k^{th} trade price on day r. Here, S_{rk} defines the time when that trade happened on the r^{th} day. Let $Q_r(S_{rk})$ be the related trading volume. The VWAP at time S_{rk} is

$$X_r(S_{rk}) := \sum_{l \le k} P_r(S_{rl})\, Q_r(S_{rl}) \Big/ \sum_{l \le k} Q_r(S_{rl}).$$

Recently, the concepts of functional data analysis have been applied to estimation of intraday volatility (e.g., Müller, Stadtmüller, and Yao, 2006; Alva, Romo, and Ruiz, 2009; Müller et al., 2011). Then, for day r, $X_r(s)$ is a log volatility process that is not observable, but can be extracted from observable quantities (Müller et al., 2011, eqn. 10).

In a functional data context, X_r is often assumed to be independent across r and it is smooth in the sense that its covariance function is continuous and possibly differentiable. For finance and economic data, the independence assumption is often violated. Smoothness of the functional data for each r can be reasonable (e.g., the option smile, the term structure of rates and the swap surface at each point in time r are smooth curves). However, differentiability can be too much of a restriction if one deals with curves generated by trajectories of semimartingales, e.g., Brownian motion. The framework of the paper is general enough to account for these situations.

When dealing with functional data, the main ingredient is an estimator of the covariance function. This is required for dimensionality reduction and to study the properties of the data in the same way we can study large dimensional data by extracting a few main factors (e.g., Dauxois, Pousse, and Romain, 1982; Bosq, 2000;

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

Ramsay and Silverman, 2005). Mutatis mutandis, estimation of the covariance function or related quantities are often needed for tests of independence, stability and group classification, regression function estimation, and variance function estimation (e.g., Ferraty, Vieu, and Viguier-Pla, 2007; Bathia, Yao, and Ziegelmann, 2010; Ferraty, 2011; Ferraty and Romain, 2011; Hu, 2013; and references therein); estimation in the frequency domain has also attracted attention recently (e.g., Panaretos and Tavakoli, 2013a, 2013b). Some of the above references also consider dependent functional data.

As an example, estimation of the covariance function is often used to compute quantities such as $\mathbb{E}_Y X_r$ for some measurable variable Y, where $\mathbb{E}_Y$ is the expectation conditional on Y. From the Karhunen–Loève expansion (e.g., Bosq, 2000), under regularity conditions,

$$\mathbb{E}_Y X_r(s) = \sum_{l=1}^{\infty} \rho_l \mathbb{E}_Y \xi_{rl} \psi_l(s),$$

where ρ_l^2 and ψ_l are the eigenvalues and eigenfunctions of the covariance operator with kernel $Cov(X_r(s), X_r(t))$ and the ξ_{rl}'s are uncorrelated real valued random variables. A consistent estimator for the covariance function is needed to compute estimators of the scores ξ_{rl}. For example, Battey and Sancetta (2013) consider estimation of the transition distribution of some positive recurrent Markov Chains setting $Y = X_{r-1}$, hence not just estimation of the first moment. Other references also look at similar or related problems where inference is based on estimators of ξ_{rl}'s (e.g., Bosq, 2000; Yao, Müller, and Wang, 2005a, 2005b; Kargin and Onatski, 2008; Müller et al., 2011).

The plan for the paper is as follows. The next section gives background material and introduces the conditions used and also provides remarks and examples around these conditions. Section 2 proposes a simple greedy algorithm for the estimation of the covariance function under the restrictions introduced by the model conditions. Section 3 addresses the consistency properties of the estimator. Consistency of the covariance function is then used to derive a Box–Pierce statistic for nonfully observable functional data. Section 4 provides details on the implementation of the estimator in practice together with a fairly complete series of simulation results to assess the relative performance of the present approach to estimation of the covariance function by local linear smoothing. It appears that there are advantages in using the present method, especially when the domain of the functional data is higher dimensional. The results of the paper are then applied to test the efficiency of the Eurodollar STIR contract on the Chicago Mercantile Exchange. Proofs are available as Electronic Supplementary Material.

1.1. The Model and Background Material

Suppose $X_r := (X(t))_{t \in \mathcal{T}}$ is a random function, with $\mathcal{T}$ a compact set. For definiteness, let $\mathcal{T} = [0,1]^K$ be the K dimensional hypercube with integer $K \geq 1$.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

Hence, for each r and t, $X_r(t)$ is a real valued random variable. Let $C_0(s,t) = Cov(X_r(s), X_r(t))$ be the covariance function of X_r, assuming stationarity with respect to $r = 1, 2, \ldots$. The covariance function is unknown and needs to be estimated. It is assumed that the covariance admits a series expansion in terms of a generalized Fourier series

$$C_0(s,t) = \sum_{\omega_1,\omega_2 \in \mathbb{N}^K} f_{\omega_1\omega_2} e_{\omega_1}(s) e_{\omega_2}(t), \tag{1}$$

where $\mathbb{N}^K$ is the K Cartesian product of the positive integers $1, 2, 3, \ldots$ (e.g., Dunford and Schwartz, 1964, IV.14). The set of functions $\mathcal{E} := \{e_\omega(s) : \omega \in \mathbb{N}^K\}$ forms an orthonormal set on $\mathcal{T}$ and are supposed to be continuous throughout the paper. In consequence, also the covariance function is continuous if for example it has finite Hilbert–Schmidt norm. Recall the Hilbert–Schmidt norm of the covariance function C_0 is $|C_0|_2 := \sqrt{\int_{\mathcal{T}} \int_{\mathcal{T}} |C_0(s,t)|^2 dsdt}$. Note that any covariance with finite Hilbert–Schmidt norm admits a representation as in (1).

Typical examples for the basis functions $\mathcal{E}$ are trigonometric or Legendre polynomials. The term generalized Fourier series is used to indicate that the expansion can be in terms of any orthonormal set of functions, not necessarily trigonometric polynomials, though continuous. For each $\omega_1, \omega_2 \in \mathbb{N}^K$, the Fourier coefficients are defined as $f_{\omega_1\omega_2} := \int_{\mathcal{T}} \int_{\mathcal{T}} C_0(s,t) e_{\omega_1}(s) e_{\omega_2}(t) dsdt$, and symmetry of C_0 in t, s implies that $\left(f_{\omega_1\omega_2} : \omega_1, \omega_2 \in \mathbb{N}^K\right)$ is symmetric: $f_{\omega_1\omega_2} = f_{\omega_2\omega_1}$. The expansion in (1) is not to be confused with Mercer's theorem, as $\mathcal{E}$ is not the set of eigenfunctions of C_0. When the Fourier basis functions are the complex exponentials, it is possible to link (1) to harmonizable and periodically correlated processes, which however differ, as they are defined over $\mathbb{R}^K$ and not $[0,1]^K$ and induce further restrictions on the covariance function (e.g., Yaglom, 1987; Dehay and Hurd, 1993; Lii and Rosenblatt, 2002).

This paper uses a sieve to approximate C_0. If any of the conditions imposed on C_0 are not valid, an approximation error is incurred and the results account for such eventuality. Hence, it is convenient to define some suitable space that can be approximated by the methods to be discussed. Consider the following subset of bilinear functions with finite Hilbert–Schmidt norm,

$$\mathcal{L}(a) := \left\{ C : C(s,t) = \sum_{\omega_1,\omega_2 \in \mathbb{N}^K} f_{\omega_1\omega_2} e_{\omega_1}(s) e_{\omega_2}(t), e_\omega \in \mathcal{E}, \sum_{\omega_1,\omega_2 \in \mathbb{N}^K} \left| f_{\omega_1\omega_2} \right| \leq a \right\}.$$

Though the notation does not make it explicit, $\mathcal{L}(a)$ does depend on $\mathcal{E}$. Denote by $\mathcal{L} := \bigcup_{a<\infty} \mathcal{L}(a)$ the union of $\mathcal{L}(a)$ for any $a < \infty$. The space $\mathcal{L}$ has interesting approximation properties under the Hilbert–Schmidt norm (Lemma 5, in Electronic Supplementary Material see also Barron, Cohen, Dahmen, and DeVore, 2008, for a discussion). Here, it is also shown that covariance functions in $\mathcal{L}(a)$ allow one to turn convergence under the Hilbert–Schmidt norm into uniform convergence (Lemma 1 in Section 3 and its byproduct for eigenfunctions: Lemma 2 in Section 3.1). This is

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

crucial when one wants to construct inferential procedures based on the scores for sparsely observed functional data (see the Box–Pierce kind of test derived in Section 3.1 for details).

1.1.1. The Conditions. Throughout, the symbol $\lesssim$ will be used to indicate inequality up to a finite absolute constant and the symbol $\asymp$ to indicate that the right and left hand sides are proportional to each other. The following conditions are used in the paper, while remarks and examples will follow right after.

Condition 1. *$(X_r)_{r\in\mathbb{N}}$ is mean zero strictly stationary beta mixing sequence of random functions (stationarity is w.r.t. r) on $\mathcal{T} = [0,1]^K$ with values in $\mathbb{R}$, with continuous covariance function C_0 on $\mathcal{T}\times\mathcal{T}$, such that $\max_{s,t}\mathbb{E}|X_1(s)X_1(t)|^\alpha < \infty$ for some $\alpha > 2$, and with mixing coefficients satisfying $\beta(k) \lesssim k^{-\beta}$ for some $\beta > (\alpha-2)/\alpha$.*

Condition 2. *The set of orthonormal basis functions $\mathcal{E}$ is continuous on $\mathcal{T}$, and $C_0 \in \mathcal{L}(a_0)$ for some $a_0 < \infty$, and for any large enough N,*

$$\sup_{\omega_1\omega_2\notin\{1,2,\ldots,N\}^K} \left|f_{\omega_1\omega_2}\right| \leq \inf_{\omega_1\omega_2\in\{1,2,\ldots,N\}^K} \left|f_{\omega_1\omega_2}\right|. \tag{2}$$

Condition 3. *$S := \left\{(S_{rk})_{k\in\{1,2,\ldots,m_r\}} : r\in\mathbb{N}\right\}$ is an independent collection of i.i.d. random variables with values in $\mathcal{T}$, independent of $(X_r)_{r\in\mathbb{N}}$, and with law absolutely continuous with respect to the Lebesgue measure. Moreover, for each r, $m_r \asymp m$, some constant that may depend on the sample size n, but $m\to\infty$ is not assumed.*

1.1.2. Discussion of the Conditions. The functional observations X_r's are strictly stationary, but allowed to be dependent across the index r. Suppose that the functional observations $(X_r(s))_{s\in\mathcal{T}}$ satisfy the Karhunen–Loève expansion

$$X_r(s) = \sum_{l=1}^{\infty} \rho_l \xi_{rl} \psi_l(s), \tag{3}$$

where $\rho_l \geq \rho_{l+1} \geq 0$, the ξ_{rl}'s are uncorrelated, $\mathbb{E}\xi_{rl} = 0$, $\mathbb{E}\xi_{rl}^2 = 1$, where the equality is in L_2. The above is true if $C_0(s,t) = \mathbb{E}X_r(s)X_r(t)$ is continuous on $\mathcal{T}\times\mathcal{T}$ (e.g., Bosq, 2000, Thm. 1.5), which is assumed throughout via Condition 2. By separability (implied by the Karhunen–Loève expansion), all the quantities are measurable (e.g., Bosq, 2000, Lem. 1.2), hence one can define mixing coefficients, in particular beta mixing coefficients. For clarity, recall the definition of beta mixing. Let $\sigma(X_r : r\leq 0)$ and $\sigma(X_r : r\geq k)$ be the sigma algebra generated by $(X_r)_{r\leq 0}$ and $(X_r)_{r\geq k}$ for $k\geq 1$. The beta mixing coefficient $\beta(k)$ for $(X_r)_{r\in\mathbb{Z}}$ is

$$\beta(k) := \mathbb{E}\sup_{A\in\sigma(X_r:r\geq k)} \left|\Pr(A|\sigma(X_r : r\leq 0)) - \Pr(A)\right| \tag{4}$$

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

(see Rio, 2000, Sect. 1.6, for other equivalent definitions). Other dependence conditions and concepts have been applied to functional data (e.g., Bosq, 2000; Hörmann and Kokoszka, 2000; Ferraty and Vieu, 2006; Kargin and Onatski, 2008; Battey and Sancetta, 2013).

In practice, the above sigma algebra is very difficult to control, consequently, the mixing coefficients cannot be derived directly. However, the sigma algebra generated by X_r coincides with the one generated by $\{\rho_l \xi_{rl} : l \geq 1\}$ (e.g., Bosq, 2000, p. 22). Hence, the functional data are beta mixing if in (4) $\sigma(X_r : r \leq 0)$ is replaced by $\sigma(\rho_l \xi_{rl} : r \leq 0, l \geq 1)$ and $\sigma(X_r : r \geq k)$ by $\sigma(\rho_l \xi_{rl} : r \geq k, l \geq 1)$. This remark can be helpful in finding the beta mixing coefficients for functional data using existing results for real valued random variables.

Example 1. In (3) suppose that the ξ_{rl}'s are independent across l, e.g., Gaussian. Let $\beta_l(k)$ be the k^{th} mixing coefficient for each score $(\rho_l \xi_{rl})_{r \in \mathbb{Z}}$. The beta mixing coefficient for independent pairs of sigma algebras is bounded by the sum of the beta mixing coefficients for each pair (e.g., Bradley, 2005, Thm. 5.1). Then, the k^{th} mixing coefficient of $(X_r)_{r \in \mathbb{Z}}$ is bounded by $\sum_{l=1}^{\infty} \beta_l(k)$, by independence of the $\rho_l \xi_{rl}$'s. This requires the scores to become independent across r as $l \to \infty$. Of course, this holds if X_r is finite dimensional, i.e., in (3) there is an $L < \infty$ such that $\rho_l = 0$ for $l \geq L$.

Beta mixing for real valued random variables is satisfied by many common models. For example, any finite order ARMA model with i.i.d. innovations and law absolutely continuous w.r.t. the Lebesgue measure satisfies absolute regularity with geometric mixing rates (Mokkadem, 1988), i.e., one can take $\beta(k) = O\left(\beta^k\right)$ for some $\beta \in (0, 1)$. Similarly, GARCH models and more generally, multivariate models that can be embedded in some stochastic recursive equation are also absolutely regular (e.g., Basrak, Davis, and Mikosch, 2002, for details: they state their results for strong mixing, but the proof shows that these models are actually absolutely regular). However, some simple processes may not satisfy absolute regularity (e.g., Andrews, 1984).

The mean zero condition on the data is used for simplicity. An estimator of the mean function does usually have a second order effect on the convergence rates (e.g., Fan and Yao, 1998; Hu, 2013).

Condition 2 is related to the regularity of the covariance function C_0. If C_0 has number of derivatives that grow with K and $\mathcal{E}$ is a set of trigonometric polynomials, or smooth wavelet basis with some vanishing moments, it can be inferred that $C_0 \in \mathcal{L}(a)$ (e.g., Barron et al., 2008). However, covariances in $\mathcal{L}(a)$ do not need to be differentiable, and if this is the case, their smoothness is more easily controlled by imposing conditions on the coefficients in the expansion. For example, let $\mathcal{E}$ be the set of trigonometric polynomials and let $K = 1$. In this case, all that is required is that the covariance function has a Fourier representation with absolutely summable coefficients. This space is more general than the usual space of twice differentiable functions considered in the literature of partially observed functional data (e.g., Conditions C in Hall, Müller, and Wang, 2006). One strand

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

of literature also studies functional data smoothed using some roughness penalty estimation method (Ramsay and Silverman, 2005). Justification of this method requires smooth trajectories, if some acceptable convergence rate is needed (see also Müller et al., 2011, for a related discussion). Recently, Cai and Yuan (2010) have considered covariance functions in a Reproducing Kernel Hilbert Space (RKHS). Often, these spaces are smooth, though, the exact meaning of this depends on the specific RKHS. In consequence, this literature often rules out examples that can be of interest in economics and finance, e.g., Brownian motion.

Example 2. Suppose that $(X_r(t))_{t\in\mathcal{T}}$ is standard Brownian motion, so that $C_0(s,t) = \min\{s,t\}$, which is not differentiable at $s = t$. It is well known that

$$C_0(s,t) = \sum_{\omega=1}^{\infty} \frac{8}{(2\omega-1)^2\pi^2} \sin\left((2\omega-1)\frac{\pi}{2}s\right)\sin\left((2\omega-1)\frac{\pi}{2}t\right)$$

(e.g., Bosq, 2000, p. 26). Then, if $e_\omega(s) = \sin\left((2\omega-1)\frac{\pi}{2}s\right)$ it is clear that Condition 2 is satisfied because $f_{\omega_1\omega_2} = \frac{8}{(2\omega-1)^2\pi^2}$ if $\omega_1 = \omega_2 = \omega$, and 0 otherwise so that the coefficients are absolutely summable.

The above example shows that the choice of basis functions can be crucial. An understanding of the econometric problem and consequent choice of basis does usually lead to a better solution than blind application of an estimation technique. However, the smoother the covariance function, the less important the choice of basis becomes. In some financial applications (e.g., the term structure of interest rates, the smile of implied volatility, the swaption surface), smoothness of the covariance function, including differentiability, is a common implicit technical condition (e.g., Duffie and Kan, 1996; Castagna and Mercurio, 2007).

Example 3. Consider an affine yield factor model of interest rates (Duffie and Kan, 1996). Then, at time r, the zero coupon bond price with expiry at $r+t$ is $X_r(t) = \exp\left\{A(t) + \sum_{l=1}^{L} B_l(t) Z_{rl}\right\}$, where $A(t)$ and the $B_l(t)$'s are smooth functions and the Z_{rl}'s are L Gaussian stochastic factors. When r is a continuous time variable the factors are diffusions, hence, Markov processes. Assume the Gaussian factors to be stationary, e.g., stationary AR(1), Ornstein–Uhlenbeck processes in continuous time or its generalizations (this is usually the case in factor models for interest rates, e.g., Brigo and Mercurio, 2006, for a review of such models). Then, by stationarity (w.r.t. to r) setting $\sigma_{kl} = Cov(Z_{rl}, Z_{rk})$, $Cov(X_r(s), X_r(t)) = \exp\left\{\sum_{l,k=1}^{L} B_l(t) B_k(s)\sigma_{kl}\right\}$ so that the smoothness of the covariance function is determined by the functions $B_l(t)$'s, which are differentiable in order to satisfy some basic regularity conditions.

Finally, (2) is a technical device to say that the basis $\mathcal{E}$ is indexed in such a way that the larger the index (in each dimension), the smaller the corresponding Fourier coefficient. This is needed to ensure that as N increases in each dimension at the same time, one can deduce the convergence rate of the approximation error.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

In some practical situations, the generalized Fourier coefficients of C_0 will satisfy

$$\left|f_{\omega_1\omega_2}\right| = O\left(\prod_k^K \left|\omega_1^{(k)}\omega_2^{(k)}\right|^{-(1+\epsilon)}\right), \tag{5}$$

for some $\epsilon > 0$, where $\omega_i^{(k)}$ is the k^{th} entry in $\omega_i := \left(\omega_i^{(1)}, \omega_i^{(2)}, \ldots, \omega_i^{(K)}\right)$, $i = 1, 2$. Then, $C_0 \in \mathcal{L}$ and (2) is satisfied. However, (2) is a weaker requirement than assuming (5). Also, one could work with sets of functions with L_2 norm equal to 1, not necessarily orthogonal (e.g., Barron, 1993; Barron et al., 2008, for a discussion). Simple examples in this case are polynomials and sigmoid functions. However, this extension requires extra technical and notational details and it is not pursued here.

Condition 3 is similar to A1.1 in Yao et al. (2005a, 2005b) and Conditions C in Hall et al. (2006). However, Hall et al. (2006) also consider the case $m \to \infty$ at a certain rate, and similarly Cai and Yuan (2010). When $m \to \infty$, improved rates are possible though never faster than the root-n rate even if nm is arbitrarily large. If the functional data are not fully observable, observation points need to be random in order to identify uniquely the covariance function, no matter how smooth the latter is. The framework is different from the local one where as the sample size increases, also the points at which we observe the functional data become more dense. This is not necessarily assumed in the partially observed functional data literature and the related applications (Yao et al., 2005a, 2005b; Hall et al., 2006; Cai and Yuan, 2010). On the other hand $m \to \infty$ is an assumption commonly used in diffusion estimation (e.g., Fan, 2005). There only a sample trajectory is available, i.e., $X_1(S_1), X_1(S_2), \ldots, X_1(S_m)$ so that m is the sample size and $n = 1$.

Independence of the observation points from the functional data is a technical simplification. One might conjecture that by suitable control of the dependence between X_r and S_k the results presented here might still hold. However, this leads to further complications and technical details.

2. ESTIMATION

Observe a sample of size n from $(X_r(S_{rk}) : k = 1, 2, \ldots, m_r)_{r\in\mathbb{N}}$. For each r, $S_{r1}, S_{r2}, \ldots, S_{rm_r}$ are the points at which X_r is observed, hence for each r, S_{rk} is a vector for $K > 1$.

Let $\theta = (\omega_1, \omega_2)$ and $\Theta_N = \{1, 2, \ldots, N\}^K \times \{1, 2, \ldots, N\}^K$. Write the truncated Fourier series representation of C_0, as $C_N(s,t) = \sum_{\theta\in\Theta_N} f_\theta \varphi_\theta(s,t)$, where

$$\varphi_\theta(s,t) = \varphi_{\omega_1,\omega_2}(s,t) = const. \times \left(e_{\omega_1}(s)\, e_{\omega_2}(t) + e_{\omega_2}(s)\, e_{\omega_1}(t)\right), \tag{6}$$

because the covariance function is symmetric in its arguments. Here, $const. = .5$ if $\omega_1 = \omega_2$, $\sqrt{.5}$ if $\omega_1 \neq \omega_2$, and it is chosen to ensure $\left|\varphi_{\omega_1,\omega_2}(s,t)\right|^2$ is normalized

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

to integrate to 1. Computationally, it is more convenient to write the above as a single summation, i.e.,

$$C_N(s,t) = \sum_{j=1}^{N(K)} f_{\theta_j} \varphi_{\theta_j}(s,t), \tag{7}$$

where $N(K)$ is the cardinality of Θ_N, and $\{\theta_1, \theta_2, \ldots, \theta_{N(K)}\} = \Theta_N$. In practice, since $\varphi_{\omega_1,\omega_2}(s,t) = \varphi_{\omega_2,\omega_1}(s,t)$ by construction, one uses $\left[N^K\left(N^K+1\right)/2\right]$ terms and the summation should run over these terms only. To avoid trivialities in the notation just suppose that $N(K) = N^{2K}$.

The researcher may estimate C_N instead of C_0 and then let $N \to \infty$. To help the researcher, suppose they do not know a_0, but they can guess an upper bound $\bar{a}$ for a_0 (a_0 as in Condition 2). Of course, if $a_0 > \bar{a}$, then, the guess is not good and a nonvanishing approximation error will be incurred, even when $N \to \infty$, unless also $\bar{a} \to \infty$. The results account for this possibility, however, to avoid extra notation it is convenient, for the moment, to assume that the guess is correct. This means that the following holds

$$\sum_{\omega_1,\omega_2 \in \mathbb{N}^K} \left| f_{\omega_1\omega_2} \right| \leq \bar{a} < \infty. \tag{8}$$

When (8) is true, one also has $\left| f_{\theta_j} \right| \leq \bar{a}$, using the equivalent notation in (7).

The following steps are suggested for the estimation algorithm. Let $J > 1$, $J \in \mathbb{N}$. Start with $j = 1$ and $\hat{C}_0 = 0$. While $j \leq J$, do the following: find

$$\left[\hat{a}_j, \hat{\theta}_j\right] = \arg \min_{a \in [-\bar{a},\bar{a}], \theta \in \Theta_N} \sum_{r=1}^{n} \sum_{k \neq l : k,l=1}^{m_r} \times \left| X_r(S_{rk}) X_r(S_{rl}) - \frac{j-1}{j} \hat{C}_{j-1}(S_{rk}, S_{rl}) - \frac{1}{j} a \varphi_\theta(S_{rk}, S_{rl}) \right|^2$$

set $\hat{C}_j(s,t) = \frac{j-1}{j} \hat{C}_{j-1}(s,t) + \frac{1}{j} \hat{a}_j \varphi_{\hat{\theta}_j}(s,t)$, and $j = j+1$.

This algorithm is a modified instance of an iterative Hilbert space approximation algorithm (Jones, 1992; Barron, 1993; Li and Barron, 2000, see Sancetta, 2013b for this constrained version). Unlike standard sieve estimation where N^{2K}/n needs to be small, here N^{2K}/n can be a large number. The precise requirements are given in Theorem 1.

The optimization at each step of the algorithm is quite fast, as it is just a regression that requires to choose the regressor $j^{-1}\varphi_\theta(s,t)$ (recall that Θ_N is a finite set) that minimizes the residual sum of the squares at each step j.

Then,

$$\hat{C}_J(s,t) = J^{-1} \sum_{j=1}^{J} \hat{a}_j \varphi_{\hat{\theta}_j}(s,t) \tag{9}$$

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

is the resulting estimator, which is a linear combination of basis functions such that $\sum_{j=1}^{J} \left|\hat{a}_j / J\right| \leq \bar{a}$. Hence, (8) is imposed in the estimation. The averaging $J^{-1} \sum_{j=1}^{J} \hat{a}_j$ allows the estimator to possess good asymptotic properties.

In particular, as $J \to \infty$, the estimator leads to the solution of the following Lasso problem:

$$\min_{f_1, f_2, \ldots, f_{N(K)}} \sum_{r=1}^{n} \sum_{k \neq l : k,l=1}^{m_r} \left| X_r(S_{rk}) X_r(S_{rl}) - \sum_{j=1}^{N(K)} f_j \varphi_{\theta_j}(S_{rk}, S_{rl}) \right|^2,$$

$$s.t. \sum_{j=1}^{N(K)} \left|f_j\right| \leq \bar{a} \tag{10}$$

(Lemma 3 in Electronic Supplementary Material). Note that the constraint above is satisfied at each iteration of the algorithm. The estimator is not guaranteed to be positive definite in a finite sample and this is common for other estimators of the covariance function, such as local linear smoothers.

The algorithm is a variation of the approximation algorithm discussed by Jones (1992) and Barron (1993) and also studied in Sancetta (2013a, 2013b). The constraint to bound the absolute value of the estimated coefficients by $\bar{a}$ is new and this is what provides direct control of the estimation error. From a practical point of view, the algorithm appears resilient to overfitting unless we allow the cardinality of Θ_N, N^{2K}, to grow exponentially fast. Essentially, this is the same conclusion reached by the L_2 boosting algorithm of Bühlmann (2006) where the absolute summability of the regression coefficients is used. Integrability conditions for the Fourier transform has been discussed by Barron (1994) to avoid the so called curse of dimensionality. Mutatis mutandis, here, this condition avoids exponential dependence on K, the dimension of $\mathcal{T}$. However, as shown by Condition 2, this requires an increasing rate of decay for the Fourier coefficients to be summable. Letting $\bar{a} \to \infty$ leads to a larger class of covariance functions dense in $\mathcal{L}$.

Section 3 provides the rates of convergence of the estimation algorithm. Computational details and numerical evidence are deferred to Section 4.

3. CONSISTENCY

THEOREM 1. *Let $\hat{C}_J$ be the estimator from the algorithm with $J > 1$ such that $J^{-1} \ln J \lesssim \sqrt{\frac{K \ln N}{n}}$. Under Conditions 1–3, if $N^K \lesssim n^p$, for some (any) $p < \infty$,*

$$\left|C_0 - \hat{C}_J\right|_2^2 = O_p\left(\bar{a}\left(1 + \frac{\bar{a} + a_0}{\sqrt{m}}\right)\sqrt{\frac{K \ln n}{n}} + \left|C_0 - C_{N,\bar{a}}\right|_2^2\right),$$

where $\hat{C}_J$ is as in (9) and $C_{N,\bar{a}}$ minimizes $|C_0 - C|_2^2$ amongst all covariance functions with representation as in (7) and such that $\sum_{\theta \in \Theta_N} |f_\theta| \leq \bar{a}$.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

Moreover, if also $n^{1/4} \lesssim N^K \lesssim n^p$ *for some* $p \in [1/4, \infty)$*, and* C_0 *satisfies (8),* $\left|C_0 - \hat{C}_J\right|_2^2 = O_p\left(\bar{a}\left(1 + \frac{\bar{a}}{\sqrt{m}}\right)\sqrt{\frac{K \ln n}{n}}\right)$.

Theorem 1 shows that the number of components N does not need to be fine tuned as is typically the case for series/sieve estimators (e.g., Example 3.4.10 in van der Vaart and Wellner, 2000). Hence, it is possible to choose N of the same order of magnitude as n, if not almost exponentially larger in order to make sure that the approximation error $\left|C_0 - C_{N,\bar{a}}\right|_2$ is small enough. Then, the complexity of the estimator is essentially determined by $\bar{a}$. This is of practical importance because for a covariance on $[0,1]^K \times [0,1]^K$, the number of terms grows exponentially in K, i.e., $N(K) = O\left(N^{2K}\right)$ so it is naturally larger than the sample size n.

Lower bounds for the rates of convergence can be used as a benchmark for the result in Theorem 1. The convergence rate from Theorem 1 is slower than the parametric rate n^{-1}. Tsybakov (2003) and Bunea et al. (2007) provide minimax rates and explicit estimators for certain function classes when the errors are Gaussian under i.i.d. sampling. When the class of functions is the convex combination of $O\left(N^{2K}\right)$ bounded functions, the convergence cannot be faster than $[(K \ln N)/n]^{1/2}$ under the expected integrated square error as soon as $N^{2K} \gtrsim \sqrt{n}$ (Tsybakov, 2003, Thm. 2). Recall that here the number of functions used is $O\left(N^{2K}\right)$ and that the convex combination of these functions is contained in the space $\mathcal{L}(\bar{a})$ when $\bar{a} \geq 1$. Hence, it is not possible to improve the rates given here, at least for fixed $\bar{a}$. In consequence, the rate of convergence is minimax when the functional data are not densely observed, i.e., m is bounded.

Covariance function estimation for densely observed functional data can achieve the parametric rate in special circumstances. This can be achieved by taking the partially observed functional observation $(X_r(S_{rk}) : k = 1, 2, \ldots, m_r)$ and approximating the values of X_r at the unobserved points by a smooth curve. For example, Hall et al. (2006) suggest using a local linear kernel smoother when X_r is not just partially observed, but also subject to additional idiosyncratic error. Then, the empirical covariance function estimator using the approximated functional data can achieve the same rate of convergence of the empirical covariance function of fully observable functional data, under regularity conditions. Amongst other things, the conditions require twice differentiable functional data, X_r to be observed at m ($m_r = m$) points with m of larger order of magnitude than $n^{1/4}$. Cai and Yuan (2010) suggest a procedure to estimate the covariance function in such a way that if the X_r's are densely observed, the rate of convergence is the parametric rate. Once again the restrictions imposed on the data are more stringent than in here.

When (8) is not satisfied, i.e., $a_0 > \bar{a}$ (a_0 as in Condition 2), one can let $\bar{a} \to \infty$. The following result also allows $a_0 \to \infty$ as long as a_0 is eventually smaller than $\bar{a}$. Hence, this generalizes the results in Theorem 1 to even more general covariance functions.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

COROLLARY 1. *Suppose that Condition 2 holds. Then,*

$$\left|C_0 - C_{N,\bar{a}}\right|_2 \lesssim a_0 N^{-K} + \max\{a_0 - \bar{a}, 0\},$$

where $C_{N,\bar{a}}$ is as in Theorem 1.

Let $\epsilon_n \to 0$ be any sequence such that $\left(\bar{a} + \frac{\bar{a}^2}{\sqrt{m}}\right)\sqrt{\frac{K \ln n}{n}} = O(\epsilon_n)$. Under Conditions 1,2, and 3, if $n^{1/4} \lesssim N^K \lesssim n^p$ for some $p \in [1/4, \infty)$, and eventually $a_0/\bar{a} \leq 1$, and $J^{-1}\bar{a}\ln(J) = O(\epsilon_n)$, then $|C_J - C_0|_2^2 = O_p(\epsilon_n)$.

The first part of Corollary 1 implicitly shows that $C_0 \in \mathcal{L}(a_0)$ imposes more stringent conditions on C_0 as the dimension K of the domain of the functional data increases. The second part shows that possibly $\bar{a}, a_0 \to \infty$, though in a restricted way. If $a_0 \to \infty$, then the "true" covariance C_0 should itself be understood as a sieve.

The convergence under the Hilbert–Schmidt norm can be turned into uniform by the following useful lemma for functions in $\mathcal{L}$.

LEMMA 1. *Let $\hat{C}_J$ be the estimator from the algorithm. Then,*

$$\max_{s,t\in\mathcal{T}}\left|C_0(s,t) - \hat{C}_J(s,t)\right| \leq \sqrt{\bar{a} + a_0}\max_{t\in\mathcal{T}}\sup_{\omega\in\mathbb{N}^K}|e_\omega(t)|^2\left|C_0 - \hat{C}_J\right|_2,$$

where a_0 is as in Condition 2.

The rate of convergence deteriorates linearly in K (compare this to the exponential growth of kernel smoothers). As shown in the Electronic Supplementary Material this is a consequence of the sparsity of the Fourier coefficients, i.e., their absolute summability. Moreover, the bound is in terms of $\bar{a}$. As shown in Corollary 1, letting $\bar{a} \to \infty$ with n allows us to consider larger classes of covariance functions. Letting $J \to \infty$ does not change the bound and it is actually recommended, as the complexity of the model is defined in terms of $\bar{a}$ and N.

Convergence in distribution for the algorithm is not possible under the conditions chosen here. In fact, the number of components N^{2K} in the approximation can be much larger than the sample size n ($N^K = O(n^p)$, for any $p < \infty$). The covariance function is predominantly needed as a quantity in some intermediary estimation step. Hence, consistency is the most important property for any such estimator.

3.1. Application: A Box–Pierce Test for Functional Data

A natural application of the estimation of the covariance function is dimensionality reduction via functional principal components. It is well known that if C_n is any estimator of the covariance function C_0, the first L estimated eigenfunctions derived from C_n, say $\{\psi_{nl} : l \leq L\}$, are close to the true eigenfunctions

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

$\{\psi_l : l \leq L\}$ of C_0, if C_n is close to C_0. In particular,

$$\int_{\mathcal{T}} |\psi_l(s) - \psi_{nl}(s)|^2 ds \leq \Delta_l^{-1} |C_0 - C_n|_2^2, \tag{11}$$

where Δ_l is proportional to the spacing between the l and $l-1$ true eigenvalue (Bosq, 2000, Lem. 4.3).

Using Lemma 1 one can also turn the above convergence into uniform. This is important when the functional data are not fully observable and one needs to control the error uniformly.

LEMMA 2. *Suppose that C_n is a covariance function such that* $\max_{s,t\in\mathcal{T}} |C_0(s,t) - C_n(s,t)| = o_p(1)$. *Suppose that C_0 is continuous on $\mathcal{T}\times\mathcal{T}$ and its first L eigenvalues are distinct and greater than zero. Then, for all $l \leq L$,* $\max_{s\in\mathcal{T}} |\psi_{ln}(s) - \psi_l(s)| = o_p(1)$, *where ψ_{ln} and ψ_l are the eigenfunctions of C_n and C_0, respectively.*

Knowledge of the eigenfunctions allows us to recover the ξ_{rl}'s in the Karhunen–Loève representation in (3), at each time r. As in principal component analysis, the l^{th} estimated score is

$$\rho_{ln}\xi_{rln} = \int_{\mathcal{T}} X_r(s)\,\psi_{ln}(s)\,ds, \tag{12}$$

where $\psi_{ln}(s)$ is the estimated eigenfunction and ρ_{ln}^2 is the l^{th} estimated eigenvalue. In practice, the integral in (12) is replaced by a sum at the points $S_{r1}, S_{r2}, \ldots, S_{rm_r}$ at which one observes $X_r(s)$. If m_r is too small, the approximation of the integral via a sum will be poor and one has to rely on other approaches (e.g., Yao et al., 2005a). For simplicity, here, it will be assumed that $m_r = m \to \infty$ with n, but no specific rate is assumed. Otherwise the analysis would require substantial departure from the standard setup. Denote the estimated scores using the approximate integral by

$$\rho_{l(n,m)}\xi_{rl(n,m)} = \frac{|\mathcal{T}|}{m}\sum_{k=1}^{m} X_r(S_{rk})\,\psi_{ln}(S_{rk}), \tag{13}$$

where $|\mathcal{T}|$ is the volume of $\mathcal{T}$. When $m \to \infty$, the above becomes a Monte Carlo integral approximation to (12) (e.g., Geweke, 1996). Of course, $|\mathcal{T}| = 1$ in the present treatment, but it is made explicit to avoid any possible confusion. Lemma 2 is a useful tool to control the error in the above approximation.

One can then test if the scores are independent. In particular, interest lies on the null hypothesis H_0: the functional data $(X_r)_{r\in\mathbb{N}}$ are i.i.d. against an alternative H_1: H_0 does not hold.

To derive the asymptotic distribution to test such hypothesis, introduce the following conditions.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

Condition 4. *$(X_r)_{r\in\mathbb{Z}}$ are independent and identically distributed, mean zero functional data with covariance function C_0 and $\max_{s,t}\mathbb{E}|X_1(s)X_1(t)|^{\alpha} < \infty$ for some $\alpha > 2$.*

Condition 5. *Fix an integer $L > 0$. Then, the first L eigenvalues $\rho_1^2, \rho_2^2, \ldots, \rho_L^2$ of C_0 are distinct and greater than zero.*

Condition 6. *The set of orthonormal basis functions $\mathcal{E}$ is continuous on $\mathcal{T}$. The estimated covariance function $C_n \in \mathcal{L}(\bar{a})$, $C_0 \in \mathcal{L}(a_0)$ for some $a_0 < \infty$, and $\sqrt{\bar{a}}\,|C_0 - C_n|_2 = o_p(1)$, i.e., $\bar{a}$ can diverge to infinity.*

Either Theorem 1 or Corollary 1 can be used to verify Condition 6. When dealing with truncated basis expansions in terms of orthonormal functions $\mathcal{E} = \{e_\omega : \omega \in \mathbb{N}\}$, the eigenfunctions can be easily found by finding the eigenvectors of the matrix of coefficients. For simplicity, consider the case $\mathcal{T} = [0,1]$ and the first N terms in $\mathcal{E}$. Construct the matrix F with the (ω_1, ω_2) entry equal to $f_{\omega_1\omega_2}$. Find the eigenvectors and eigenvalues of F. Assume the eigenvalues are sorted in decreasing order. Let ρ_l^2 be the l^{th} eigenvalue and denote by $p_{l\omega}$ the ω^{th} entry in the corresponding eigenvector. Then, the l^{th} eigenfunction is equal to $\psi_l(s) = \sum_{\omega\le N} p_{l\omega}e_\omega(s)$. The construction requires an orthonormal basis indexed in the natural numbers. Hence, the extension to $\mathcal{T} = [0,1]^K$ with $K > 1$ is just a matter of reindexing the basis $\mathcal{E}$ which has elements indexed in $\mathbb{N}^K$ rather than $\mathbb{N}$. Consider $\left\{e_\omega : \omega \in \{1,2,\ldots,N\}^K\right\}$, i.e., the first N indexes in each dimension. Then, redefine the index $\{1,2,\ldots,N\}^K$ as $\{1,2,3,\ldots,N,N+1, N+2,\ldots,N^K\}$, as there are N^K terms, and $\tilde{e}_j = e_\omega$ with $j = \omega^{(1)} + \sum_{k=2}^{K}\left(\omega^{(k)} - 1\right)N^{k-1}$, where $\omega^{(k)}$ is the k^{th} entry in ω. The above display maps uniquely an element ω in $\{1,2,\ldots,N\}^K$ to an element j in $\left\{1,2,\ldots,N^K\right\}$. The specific mapping is actually irrelevant, and any one-to-one correspondence would work. Since $\left\{\tilde{e}_j : j \in \left\{1,2,\ldots,N^K\right\}\right\}$ is an orthonormal basis, then one can construct the matrix $\tilde{F}$ with (i,j) entry $\tilde{f}_{ij}$ and find its eigenvalues and eigenvectors with $\tilde{p_{lj}}$ being the j entry in the l^{th} eigenvector. The l^{th} eigenfunction is then $\psi_l(s) = \sum_{j\le N^K} p_{lj}\tilde{e}_j(s)$.

The result below derives the asymptotic distribution to test the null hypothesis.

THEOREM 2. *Suppose Conditions 3–6 hold. Define*

$$\gamma_{(n,m)}(k,l) = \frac{1}{n}\sum_{r=1}^{n-k}\left(\rho_{l(n,m)}\xi_{rl(n,m)}\right)\left(\rho_{l(n,m)}\xi_{r+k,l(n,m)}\right),$$

with $\rho_{l(n,m)}\xi_{rl(n,m)}$ as in (13). Then, for any fixed finite integers $Q, L > 0$, the Box–Pierce statistic

$$n\sum_{l=1}^{L}\sum_{k=1}^{Q}\left(\frac{\gamma_{(n,m)}(k,l)}{\gamma_{(n,m)}(0,l)}\right)^2 \to \chi^2_{(QL)}$$

in distribution if $m \to \infty$ with n.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

Above, Q is the number of lags to include in the test, while L denotes the number of factor scores used in the test. Mutatis mutandis, the same result holds for the Ljung–Box test. The usefulness of this test relies on the fact that we are actually testing independence of the functional data across r using only a few variables. Note that any consistent covariance function is allowed as long as Condition 6 is satisfied. However, the convergence rate of the covariance controls the convergence of the eigenfunctions (e.g., (11)). This rate of convergence affects the rate of weak convergence in Theorem 2.

Horváth and Kokoszka (2012, Thm. 2, Chap. 7) state a similar result. However, the framework there is the one of fully observable functional data. That is equivalent to have the r.h.s. term $o_p(1)$ in Condition 6 replaced by $O_p\left(n^{-1/2}\right)$ and being able to compute (12) directly. The asymptotics for fully observable functional data follows from probabilistic results of random variables in Banach space. Hence, parametric rates are possible in that case. Here, one is not even making the assumption that (12) can be computed exactly, as it is often the case in practice.

4. NUMERICAL EXAMPLES

For the sake of clarity, the minimization of

$$\sum_{r=1}^{n}\sum_{k\neq l=1}^{m_r}\left|X_r(S_{rk})X_r(S_{rl})-\frac{j-1}{j}\hat{C}_{j-1}(S_{rk},S_{rl})-\frac{1}{j}a\varphi_\theta(S_{rk},S_{rl})\right|^2 \tag{14}$$

requires to find for each $\upsilon \in \left\{1,2,\ldots,N^{2K}\right\}$ (recall we look over N^K integer values for ω_1 and ω_2)

$$a_\upsilon=\frac{j\sum_{r=1}^{n}\sum_{k\neq l:k,l=1}^{m_r}\varphi_{\theta_\upsilon}(S_{rk},S_{rl})\left(X_r(S_{rk})X_r(S_{rl})-\frac{j-1}{j}\hat{C}_{j-1}(S_{rk},S_{rl})\right)}{\sum_{r=1}^{n}\sum_{k\neq l:k,l=1}^{m_r}\varphi^2_{\theta_\upsilon}(S_{rk},S_{rl})}$$

and to project it on $[-\bar{a},\bar{a}]$, i.e., $a_\upsilon=\max\left(\min\left(a_\upsilon,\bar{a}\right),-\bar{a}\right)$ and then choose the a_υ and the corresponding θ_υ that minimize (14). Using R or Matlab this procedure can be vectorized and requires to run $O\left(N^{2K}\right)$ univariate regressions. For the problems discussed here, running the algorithm only required a few seconds.

In the numerical examples to follow, the tensor product of shifted Legendre polynomials on $[0,1]^K$ was used for $\varphi_{\omega_1,\omega_2}(s,t)$, with the latter as in (6). In particular $e_{\omega_1}(s)=\prod_{k=1}^{K}e^{(k)}_{\omega_1}\left(s^{(k)}\right)$, where $e_{\omega_1^{(k)}}(s)$ is the shifted Legendre polynomial of order $\omega_1^{(k)}-1$ on $[0,1]$. As usual the superscript (k) denotes the k^{th} element in a K dimensional vector. For example, for $\omega_1^{(k)}=1,2$, $e_1\left(s^{(k)}\right)=1$, $e_2\left(s^{(k)}\right)=\sqrt{2}\left(2s^{(k)}-1\right)$. This means taking a Legendre polynomial, which is defined on $[-1,1]$ and using the transform $x^{(k)}=2s^{(k)}-1$ (hence the name shifted) and standardizing to ensure its square integrates to one.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

Most of the Matlab code used for the estimation is available from the author's webpage and it includes the tensor product Legendre basis generation, the vectorized implementation of the algorithm, and the calculation of the estimated scores and the test statistics from Theorem 2.

4.1. Monte Carlo Experiment

The estimator is compared to the usual bivariate local linear estimator of the covariance function. As done elsewhere in a functional data context (e.g., Cai and Yuan, 2010; Battey and Sancetta, 2013), data are generated according to a random model. This reduces the dependence of the results on the Monte Carlo design. In particular, for $K = 1$,

$$X_r\left(S_{k_r}\right) = \sum_{\omega=1}^{100} \left(\frac{a_\omega}{\omega^{1.1}}\right) Z_{r\omega} \cos\left(\omega\pi\, S_{k_r}\right), \tag{15}$$

where the $Z_{r\omega}$'s are standard normal, i.i.d. across r, and have correlation .75 across ω. The above does not imply that the eigenfunctions are cosines because the $Z_{r\omega}$ are correlated. In fact, the covariance function is equal to

$$C_0(s,t) = \sum_{\omega_1=1}^{100}\sum_{\omega_2=1}^{100} 0.75 \frac{a_{\omega_1} a_{\omega_2}}{(\omega_1\omega_2)^{1.1}} \cos(\omega_1\pi s)\cos(\omega_2\pi t). \tag{16}$$

For each simulation, a sample of $n = 10, 50$ observations is generated from the above model where the constants a_ω's are random samples from a $[0,1]$ uniform distribution while for each r, $\{S_{rk} : k = 1, 2, \ldots, m\}$ are $[0,1]$ uniform with $m = 4, 20$. The total number of simulations is 250. This framework is quite general.

While C_0 admits a trigonometric expansion, the basis used in the estimation will be shifted Legendre polynomials up to order 9 plus a constant, i.e., 10 elements. Exploiting symmetry, this means that the total number of possible coefficients to estimate equals $10 \times 11/2 = 55$. Of course, the estimation procedure only leads to a sparse estimator as in Lasso (see end of Section 2) so only a handful of coefficients are nonzero (e.g., 5).

The case $K = 2$ is also considered. Then, the 2 dimensional half period cosines are constructed via tensor product. To ease the computational burden, in this case, the $Z_{r\omega}$'s are uncorrelated across ω. This means that the eigenfunctions are product of cosines. The S_{rk}'s are now $[0,1]^2$ uniform random variables. The order of the Legendre polynomials in each dimension is chosen to be 4 plus the constant, i.e., 5 elements in each dimension. This means that the total number of possible coefficients to estimate is $5^2 \times \left(5^2+1\right)/2 = 325$.

The current estimator under semiparametric restriction (SP) was computed with $J = 1,000/K$ (division by K is for convenience when $K = 2$) and $\bar{a} \in A \times \{1, 3, 9, 27, 81\}$ chosen by cross-validation, where $A = \frac{1}{nm}\sum_{r=1}^{n}\sum_{k_r=1}^{m}\left|X\left(S_{k_r}\right)\right|$. For comparison the covariance was also estimated by a local linear smoother.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

The smoothing matrix was chosen to be hI_K (I_K the K dimensional identity), and $h = c\frac{\sqrt{\sigma_{XX}\sigma_S}}{0.6745}\left(\frac{4}{n(K+2)}\right)^{\frac{1}{K+4}}$, where σ_{XX} and σ_S are in sample mean absolute deviations for $[X_1(S_1)\, X_1(S_2)]$ and S_1, respectively. Hence, the above is one of the several rules of thumb for bandwidth selection. The constant $c \in \{.5, 1, 1.5, 2, 3\}$ was chosen by cross-validation for each iteration. The simulation results showed that the SP estimator provides good performance against the local linear smoother. Of course, a different Monte Carlo setup could reverse the picture. Due to space constraints, only results for $n = 10$, $m = 4$, and $K = 1, 2$ are reported. Figure 1,

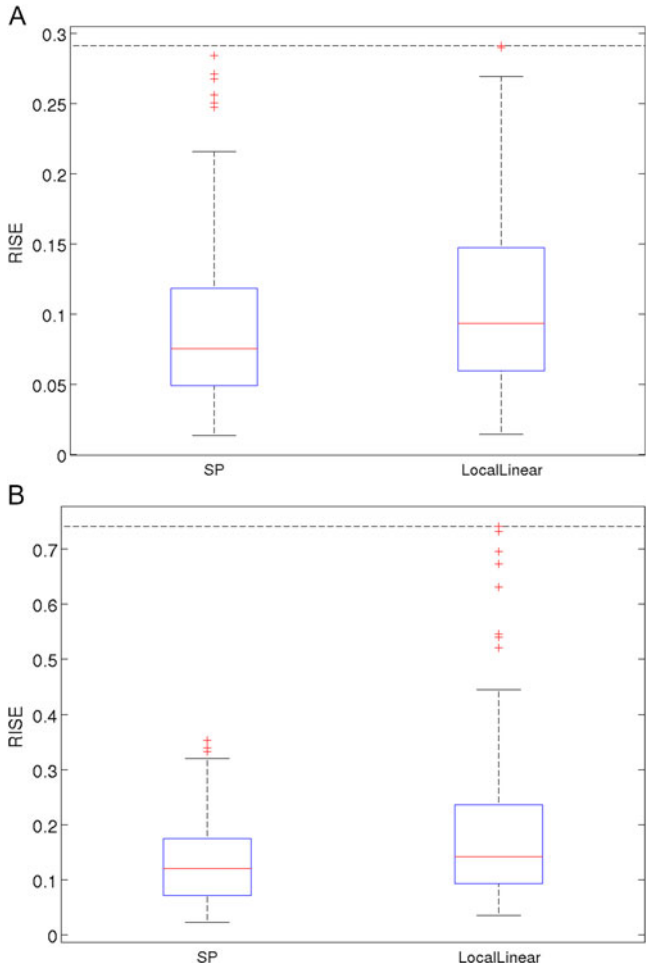


FIGURE 1. Box plot for SP and local-linear smoother. Panel A: $n = 10$, $m = 4$, $K = 1$. Panel B: $n = 10$, $m = 4$, $K = 2$.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

Panels A and B, reports the box plots for the root integrated square error (RISE). Results for other box plots are qualitatively the same and reported in the Electronic Supplementary Material.

4.2. Testing Efficiency of Term Structure of Eurodollar Futures Contracts

Consider the term structure of interest rates implied by the Eurodollar futures contracts traded on the Chicago Mercantile Exchange (CME). The contracts underlying is the 3-month Euribor. The contracts trade for different expiry dates and details can be found on the CME URL. The data used here are the smoothed curves used in Kargin and Onatski (2008, Sect. 3) and available from http://www.prismnet.com/˜slava/papers/ED_rates_CRB.txt. Given that the curves can be considered to be fully observed, the framework is within the scope of the paper by letting m be very large.

Each observation is a vector of yields describing the yield curve for that day, where each point on the curve is a tenor 30 days apart from the previous one. Hence, the curves are observed at equally spaced points. Each vector has size 114, i.e., 114 months. CME Eurodollar contracts are priced such that if P_r is the price of a contract on day r, $1 - P_r$ is the annualized yield. If the prices are efficient, the price of each contract behaves as a random walk, and the contract prices are integrated of order 1. Hence by first differencing and changing the sign, define $X_r = P_{r-1} - P_r$ to be the change in yield. The observation points on each day are equally spaced and standardized to lie in $[0, 1]$, e.g., $S_{rk} = k/114$, $k = 1, 2, \ldots, 114$ for all r's, so one can drop the r subscript. Hence, by construction, $1 - P_r(S_k)$ is the $114 \times S_k$-month forward on the 3 months Euribor yield on day r. In turn this implies that $X_r(S_k)$ is the daily change in yield for a 3 months deposit which starts to mature in $114 \times S_k$ months. After differencing, the data have been demeaned using the sample mean over the period considered. In particular, 3 different periods of 3 months each are considered: 01/Mar/1999–28/May/1999, 01/Mar/2001–31/May/2001, 01/Jul/2003–30/Sep/2003. Longer periods may lead to spurious conclusions due to some forms of heterogeneity. The periods correspond to the Brazilian crisis, the after dot-com bubble with the Fed reducing rates from 5.5% to 4% over those three months, and, lastly, the period when the Fed lowered rates to 1% and kept it there for one year. The demeaned functional data for the third period are plotted in Figure 2, Panel A. Due to liquidity issues, contracts with expiry after 24–36 months tend to be less actively traded. This is reflected in the yield differences being flat after 24–36 months. Figures showing the data trajectories and eigenfunctions for the other two periods are available in the Electronic Supplementary Material.

One could have considered the discretized data directly and conducted a test of independence on all the points across time. In this case, one would be dealing with a multidimensional problem with 114 variables. Without the functional data framework such a problem would require a very large sample size. In fact, one

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

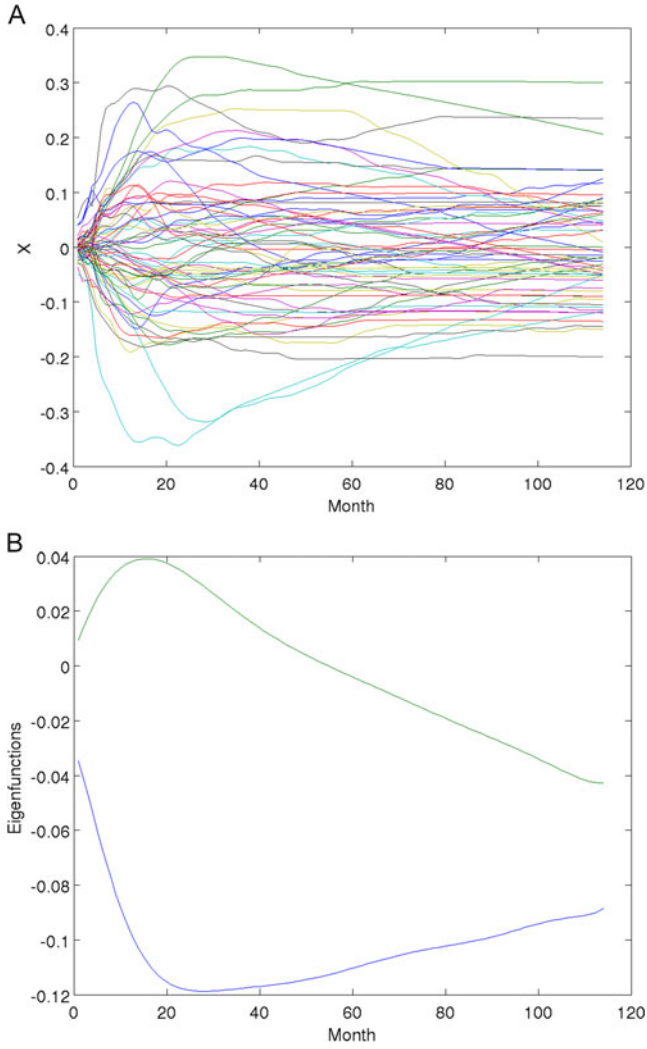


FIGURE 2. Functional data during 01/Jul/2003–30/Sep/2003. Panel A: centered data. Panel B: first 2 eigenfunctions.

would be required to estimate the covariance between 114 variables with their lags, for a total of $O\left(114^2 \times L\right)$ univariate sample covariances.

The forward yield curve can be first decomposed into a few factors that explain a large proportion of the variation, say 95%. For the present dataset this leads to 2 components. This is in line with other studies of the term structure: usually, only up to 3 components are retained, commonly termed level, slope, and curvature (e.g., Diebold and Li, 2006). These factors can also be related to tradeable

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

instruments, i.e., futures on the level, calendar spreads, and butterflies. Given that the data have been first differenced and the sample mean subtracted from it, one expects the first factor, i.e., the constant level, to drop from the sample calculations.

The eigenfunctions are computed using a shifted Legendre polynomial basis. Because of the aforementioned remarks about level, slope, and curvature, the choice seems appropriate. The eigenfunctions for the three different periods are similar. Figure 2, Panel B, shows the first two eigenfunctions for the third period, where the lower line is the first eigenfunction and the upper one is the second eigenfunction. The eigenfunctions seem to capture the salient features that emerge from looking at the data in Figure 2, Panel A. There is increasing variability as one approaches the 20^{th} month and then it progressively flattens out. Given the aforementioned remarks on taking first differences and subtracting the sample mean, the first two eigenfunctions should mostly reflect changes in slope and curvature, which seems to be the case up to the 20^{th} month.

To test the hypothesis of independence of the yield changes, the results in Section 3.1 are used with 5 lags. It is found that the null of independence is only rejected for the subsample corresponding to the after dot-com bubble. Hence it is possible to infer some level of predictability in "slope" and "curvature". This is in line with the anecdotal claims by traders and practitioners that trading opportunities increase during crisis. The actual results are reported as a table in the Electronic Supplementary Material, where implementation details are also included.

REFERENCES

Almgren, R. (2003) Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance* 10, 1–18.

Almgren, R. & N. Chriss (2000) Optimal execution of portfolio transactions. *Journal of Risk* 3, 5–39.

Alva, K., J. Romo, & E. Ruiz (2009) Modelling Intra-Daily Volatility by Functional Data Analysis: An Empirical Application to the Spanish Stock Market. Working paper 09-28, Statistics and Econometrics, Departamento de Estadística, Universidad Carlos III de Madrid.

Andrews, D.W.K. (1984) Nonstrong mixing autoregressive processes. *Journal of Applied Probability* 21, 930–934.

Barron, A.R. (1993) Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* 39, 930–944.

Barron, A.R. (1994) Approximation and estimation bounds for artificial neural networks. *Machine Learning* 14, 113–143.

Barron, A.R., A. Cohen, W. Dahmen, & R.A. DeVore (2008) Approximation and learning by greedy algorithms. *Annals of Statistics* 36, 64–94.

Basrak, B., R.A. Davis, & T. Mikosch (2002) Regular variation of GARCH processes. *Stochastic Processes and Their Applications* 99, 95–115.

Bathia, N., Q. Yao, & F. Ziegelmann (2010) Identifying the finite dimensionality of curve time series. *Annals of Statistics* 38, 3352–3386.

Battey, H. & A. Sancetta (2013) Conditional estimation for dependent functional data. *Journal of Multivariate Analysis* 120, 1–17.

Bisesti, L., A. Castagna, & F. Mercurio (2005) Consistent pricing and hedging of an FX options book. *Kyoto Economic Review* 74, 65–83.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

Bosq, D. (2000) *Linear Processes in Function Spaces: Theory and Applications*. Lecture Notes in Statistics, 149. Springer.

Boucheron, S., G. Lugosi and P. Massart (2003) Concentration Inequalities Using the Entropy Method. Annals of Probability 31, 1583-1614.

Bradley, R.C. (2005) Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys* 2, 107–144.

Brigo, D. & F. Mercurio (2006) *Interest Rate Models – Theory and Practice*. Springer.

Bugni, F.A. (2012) Specification test for missing functional data. *Econometric Theory* 28, 959–1002.

Bugni, F.A., P. Hall, & J.L. Horowitz (2009) Goodness-of-fit tests for functional data. *The Econometrics Journal* 12(Tenth Anniversary Issue), S1–S18.

Bühlmann, P. (2006) Boosting for high-dimensional linear models. *Annals of Statistics* 34, 559–583.

Bunea, F., A. Tsybakov and M. Wegkamp (2007) Aggregation for Gaussian Regression. Annals of Statistics 35, 1674-1697.

Cai, T. and M. Yuan (2010) Nonparametric Covariance Function Estimation for Functional and Longitudinal Data. Technical Report.

Castagna, A. & F. Mercurio (2007) The Vanna–Volga method for implied volatilities. *Risk Magazine*, 106–111.

Dauxois, J., A. Pousse, & Y. Romain (1982) Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis* 12, 136–154.

Dehay, D. & H.L. Hurd (1993) Representation and estimation for periodically and almost periodically correlated random processes. In W.A. Gardner (ed.), *Cyclostationarity in Communications and Signal Processing*. IEEE Press.

DeVore, R. & V. Temlyakov (1996) Some remarks on greedy algorithms. *Advances in Computational Mathematics* 5, 173–187.

Diebold, F.X. & C. Li (2006) Forecasting the term structure of government bond yields. *Journal of Econometrics* 130, 337–364.

Doukhan, P., P. Massart, & E. Rio (1994) The functional central limit theorem for weakly dependent processes. *Annales de l'institut Henri Poincarè, Probabilitès et Statistiques* 30, 63–82.

Duffie, D. & R. Kan (1996) A yield-factor model of interest rates. *Mathematical Finance* 6, 379–406.

Dunford, N. & J.T. Schwartz (1964) *Linear Operators*, vol. 1. Interscience Publishers.

Fan, J. (2005) A Selective Overview of Nonparametric Methods in Financial Econometrics (with discussion). Statistical Science 20, 317-357.

Fan, J. & Q. Yao (1998) Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* 85, 645–660.

Ferraty, F. & P. Vieu (2006) *Nonparametric Functional Data Analysis*. Springer.

Ferraty, F., P. Vieu, & S. Viguier-Pla (2007) Factor-based comparison of groups of curves. *Computational Statistics & Data Analysis* 51, 4903–4910.

Ferraty, F., A. Quintela-del-Río, & P. Vieu (2012) Specification test for conditional distribution with functional data. *Econometric Theory* 28, 363–386.

Ferraty, F. (ed.) (2011) *Recent Advances in Functional Data Analysis and Related Topics*. Springer.

Ferraty, F. & Y. Romain (eds.) (2011) *The Oxford Handbook of Functional Data Analysis*. Oxford University Press.

Geweke, J. (1996) Monte Carlo simulation and numerical integration. In H.M. Amman, D.A. Kendrick, & J. Rust (eds.), *Handbook of Computational Economics*, vol. 1, pp. 731–800. Elsevier.

Hall, P., H.G. Müller, & J.-L. Wang (2006) Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics* 34, 1493–1517.

Hasbrouck, J. (2007) *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press.

Hörmann, S. & P. Kokoszka (2000) Weakly dependent functional data. *Annals of Statistics* 38, 1845–1884.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press

Horváth, L. & P. Kokoszka (2012) *Inference for Functional Data with Applications*. Springer.
Hu, Y. (2013) Nonparametric estimation of variance function for functional data under mixing conditions. *Communications in Statistics – Theory and Methods* 42, 1774–1786.
Jones, L.K. (1992) A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics* 20, 608–613.
Kargin, V. & A. Onatski (2008) Curve forecasting by functional autoregression. *Journal of Multivariate Analysis* 99, 2508–2526.
Li, J.Q. & A.R. Barron (2000) Mixture density estimation. In S.A. Solla, T.K. Leen, & K.-R. Müller (eds.), *Advances in Neural Information Processing Systems*, vol. 12, pp. 279–285. MIT Press.
Lii, K.-S. & M. Rosenblatt (2002) Spectral analysis for harmonizable processes. *Annals of Statistics* 30, 258–297.
Mokkadem, A. (1988) Mixing properties of ARMA processes. *Stochastic Processes and Their Applications* 29, 309–315.
Müller, H.G., S. Rituparna, & U. Stadtmüller (2011) Functional data analysis for volatility. *Journal of Econometrics* 165, 233–245.
Müller, H.G., U. Stadtmüller, & F. Yao (2006) Functional variance processes. *Journal of American Statistical Association* 101, 1007–1018.
Panaretos, V.M. & S. Tavakoli (2013a) Cramèr–Karhunen–Loève representation and harmonic principal component analysis of functional time series. *Stochastic Processes and Their Applications* 123, 2779–2807.
Panaretos, V.M. & S. Tavakoli (2013b) Fourier analysis of stationary time series in function space. *Annals of Statistics* 41, 568–603.
Pollard, D. (2002) Maximal inequalities via bracketing with adaptive truncation. *Annales de l'institut Henri Poincarè, Probabilitès et Statistiques* 38, 1039–1052.
Ramsay, J.O. & J.B. Ramsey (2002) Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Econometrics* 107, 327–344.
Ramsay, J.O. & B.W. Silverman (2005) *Functional Data Analysis*, 2nd ed. Springer.
Rio, E. (2000) *Thèorie Asymptotique des Processus Alèatoires Faiblement Dèpendants*. Springer.
Sancetta, A. (2013a) A recursive algorithm for mixture of densities estimation. *IEEE Transactions on Information Theory* 59, 6893–6906.
Sancetta, A. (2013b) Greedy algorithms for prediction. Preprint, URL: https://sites.google.com/site/wwwsancetta/academicpublications/greedy-algorithms-for-prediction-1.
Tsybakov, A.B. (2003) Optimal Rates of Aggregation. Proceedings of COLT-2003, Lecture Notes in Artificial Intelligence, 303-313.
Van der Vaart, A. & J.A. Wellner (2000) *Weak Convergence and Empirical Processes*. Springer.
Yao, F., H.-G. Müller, & J.-L. Wang (2005a) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100, 577–590.
Yao, F., H.-G. Müller, & J.-L. Wang (2005b) Functional linear regression analysis for longitudinal data. *Annals of Statistics* 33, 2873–2903.

https://doi.org/10.1017/S0266466614000784 Published online by Cambridge University Press