The Annals of Statistics
2019, Vol. 47, No. 1, 1–38
https://doi.org/10.1214/17-AOS1590

FUNCTIONAL DATA ANALYSIS BY MATRIX COMPLETION[1]

BY MARIE-HÉLÈNE DESCARY AND VICTOR M. PANARETOS

Ecole Polytechnique Fédérale de Lausanne

Functional data analyses typically proceed by smoothing, followed by functional PCA. This paradigm implicitly assumes that rough variation is due to nuisance noise. Nevertheless, relevant functional features such as time-localised or short scale fluctuations may indeed be rough relative to the global scale, but still smooth at shorter scales. These may be confounded with the global smooth components of variation by the smoothing and PCA, potentially distorting the parsimony and interpretability of the analysis. The goal of this paper is to investigate how both smooth and rough variations can be recovered on the basis of discretely observed functional data. Assuming that a functional datum arises as the sum of two uncorrelated components, one smooth and one rough, we develop identifiability conditions for the recovery of the two corresponding covariance operators. The key insight is that they should possess complementary forms of parsimony: one smooth and finite rank (large scale), and the other banded and potentially infinite rank (small scale). Our conditions elucidate the precise interplay between rank, bandwidth and grid resolution. Under these conditions, we show that the recovery problem is equivalent to rank-constrained matrix completion, and exploit this to construct estimators of the two covariances, without assuming knowledge of the true bandwidth or rank; we study their asymptotic behaviour, and then use them to recover the smooth and rough components of each functional datum by best linear prediction. As a result, we effectively produce separate functional PCAs for smooth and rough variation.

1. Introduction. Functional principal component analysis, the empirical version of the celebrated Karhunen–Loève expansion, is arguably the workhorse of Functional Data Analysis (Bosq [2], Ramsay and Silverman [18], Horvath and Kokoszka [10], Hsing and Eubank [11], Wang et al. [20]). It aims to construct a parsimonious yet accurate finite dimensional representation of n observable i.i.d. replicates $\{X_1, \dots, X_n\}$ of a real-valued random function $\{X(t) : t \in [0, 1]\}$ under study. The sought representation is in terms of a Fourier series built using the eigenfunctions $\{\varphi_k\}$ of the integral operator $\mathscr{R}$ with kernel $\mathrm{Cov}(X(t), X(s))$. Such a finite-dimensional representation is key in functional data analysis: not only does it serve as a basis for motivating methodology by analogy to multivariate statistics,

Received August 2016; revised February 2017.

[1] Research supported by a Swiss National Science Foundation grant.

MSC2010 subject classifications. Primary 62M; secondary 15A99, 62M15, 60G17.

Key words and phrases. Analyticity, banding, covariance operator, functional PCA, low rank, resolution, scale, smoothing.

but it constitutes the canonical means of regularization in regression, testing, and prediction, which are all ill-posed inverse problems when dealing with functional data; see Panaretos and Tavakoli [17] for an account of the genesis and evolution of functional PCA and Wang et al. [20] for an overview of its manifold applications in functional data analysis.

Since the covariance operator $\mathscr{R}$ is unknown in practice, functional PCA must be based on its empirical counterpart (Dauxois et al. [7], Bosq [2]),

$$\hat{\mathscr{R}}_n = \sum_{i=1}^{n} (X_i - \overline{X}) \otimes (X_i - \overline{X}) \qquad \text{where } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Even this, however, is seldom accessible: one cannot perfectly observe the complete sample paths of $\{X_1, \dots, X_n\}$. Instead, one has to make do with discrete measurements

$$X_{ij} = X_i(t_j) + \varepsilon_{ij}, \qquad i = 1, \dots, n, j = 1, \dots, K, \tag{1.1}$$

where the points t_j can be random or deterministic and the array ε_{ij} is assumed to be comprised of centred i.i.d. perturbations, independent of the X_i (see, e.g., Ramsay and Silverman [18], Hall et al. [9], Li and Hsing [14]). Roughly speaking, there are two major approaches to deal with discrete measurements: to smooth the discretely observed curves and then obtain the covariance operator and spectrum of the smooth curves; and the converse, that is, to first obtain a smoothed estimate of the covariance operator and to use this to estimate the unobservable curves and their spectrum.

The first general approach was popularised by Ramsay and Silverman [18], by means of smoothing splines, and is widely used, chiefly when the observation grid $\{t_1, \dots, t_K\}$ is sufficiently dense. One defines smoothed curves $\widetilde{X}_i$ as

$$\widetilde{X}_i(t) = \arg \min_{f \in C^2[0,1]} \left\{ \sum_{j=1}^{K} \left(f(t_j) - X_{ij}\right)^2 + \tau \|\partial_t^2 f\|_{L^2}^2 \right\}, \qquad i = 1, \dots, n, \tag{1.2}$$

for $C^2[0, 1]$ the space of twice continuously differentiable functions on [0, 1], and $\tau > 0$ a regularising constant. The proxy curves $\{\widetilde{X}_i\}$ are used in lieu of the unobservable $\{X_i\}$ in order to construct a "smooth" empirical covariance operator $\widetilde{\mathscr{R}}$, and the curves $\{\widetilde{X}_i\}$ are finally projected onto the span of the first r eigenfunctions of $\widetilde{\mathscr{R}}$.

A second general approach, Principal Analysis by Conditional Expectation (PACE), was introduced by Yao et al. [21] (see also Yao et al. [22]), motivated by the need to consider situations where the grid is sparse and curves are sampled at varying grid points. In our sampling setup, and assuming the array $\{\varepsilon_{ij}\}$ to be i.i.d. of variance σ^2, they exploit the fact that the $K \times K$ covariance matrix of the vector $(X_{i1}, \dots, X_{iK})^\top$ equals (up to a factor) $\rho(t_i, t_j) + \sigma^2 \mathbf{1}\{i = j\}$. Thus, the effect of the term ε is restricted to the addition of a σ^2-ridge to the diagonal.

Yao et al. [21] then delete the diagonal $i = j$ of the empirical covariance matrix of $\{X_{ij}; i = 1, \dots, n; j = 1, \dots, K\}$ and smooth what remains to obtain a smooth estimate $\widetilde{\rho}(s,t)$ of the kernel $\rho(s,t)$. The smoothing assumes (and induces) C^2-level behaviour near $t = s$. The kernel $\widetilde{\rho}(s,t)$ is then used to construct mean-square optimal predictors $\{\widetilde{X}_1, \dots, \widetilde{X}_n\}$ of the unobservable sample paths, truncated to belong to the span of the first r eigenfunctions of $\tilde{\rho}(s,t)$.

Proceeding in either of these two ways essentially consigns any variations of smoothness class less than C^2 to pure noise, and subsequently smears them by means of smoothing; any further rough variations are expected to be negligible, and due to small fluctuations around eigenfunctions of order at least $r + 1$ (thus orthogonal to the smooth variations) and are also discarded post-PCA.

Mathematically speaking, "smooth-then-PCA" approaches correspond to an underlying ansatz that $X(t)$ is well approximated by the sum of two uncorrelated components: a "true signal" $Y(t)$ of (essentially) finite rank r and of smoothness class C^k $(k \geq 2)$ and a noise component $W(t)$ whose covariance kernel is a scaled delta function $\sigma^2\delta(s-t)$, corresponding to white noise:

$$X_i(t) = Y_i(t) + W_i(t), \qquad i = 1, \dots, n, \tag{1.3}$$

$$X_{ij} = Y_i(t_j) + W_i(t_j) = Y_i(t_j) + \varepsilon_{ij}, \qquad i = 1, \dots, n; j = 1, \dots, K. \tag{1.4}$$

The first equation can formally be understood only in the weak sense as an SDE, and in reality W would have a covariance supported on some band $\{|t - s| < \delta\}$ for some infinitesimally small $\delta > 0$. The construction of the rank r version (by PCA) of the smoothed curves $\{\widetilde{X}_i(t)\}$ can thus be seen as an the estimation of the unobservable $\{Y_i(t)\}$. Any residual variation is then indirectly attributed to W_i, seen as functional residuals, and subsequently ignored.

It may very well happen, though, that W be rough but still be mean-square continuous, possessing a covariance kernel $b(s,t) = b(s,t)\mathbf{1}\{|t-s| < \delta\}$, for b a continuous nonconstant function and $\delta > 0$ nonnegligible: "*the functional variation that we choose to ignore is itself probably smooth at a finer scale of resolution*" (Ramsay and Silverman [18], Section 3.2.4). In this case, the rough variations are not due to pure noise, but to actual signal, and contain second-order structure that we may not wish to confound with that of Y or discard. Quite to the contrary, it should be fair game for functional data analysis to aim to deal with variations at smaller scales δ; to quote Ramsay and Silverman [18], Section 3.2.4, again: "*this can pay off in terms of better estimation, and this type of structure may be in itself interesting; a thoughtful application of functional data analysis will always be open to these possibilities*". To accommodate a nontrivial kernel $b(s,t)$, the smoothing spline approach would need to replace the "uncorrelated" objective function in equation (1.2), with the "correlated" version

$$\widetilde{X}_i(t) = \underset{f \in C^2[0,1]}{\arg\min} \{(\mathbf{X}_i - \mathbf{f})B^{-1}(\mathbf{X}_i - \mathbf{f})^\top + \tau \|\partial_t^2 f\|_{L^2}^2\}, \tag{1.5}$$

for B the covariance matrix of $(W_i(t_1), \dots, W_i(t_K))^\top$, $\mathbf{X}_i = (X_{i1}, \dots, X_{iK})^\top$ and $\mathbf{f} = (f(t_1), \dots, f(t_K))^\top$. Unfortunately, B is unknown, and worse still, B and $X_i(t)$ are not jointly identifiable without further (parametric) restrictions (see Opsomer et al. [16]). Similarly, the PACE approach would need to remove a nontrivial band around the diagonal of the empirical covariance operator prior to smoothing; this would lead to unidentifiability without further assumptions. It would seem that the two approaches cannot be remedied by means of a simple modification, and a novel approach would be needed.

The aim of the paper is to put forward such a novel approach and to fill this gap. Without assuming knowledge of the rank r or the scale δ, we set out to:

1. Determine nonparametric conditions under which the smooth and rough variation are jointly identifiable on the basis of discrete data, and elucidate how the effective rank r of the smooth component, the scale δ of the rough component, and the grid resolution K affect identifiability.
2. Construct estimators of the covariance structure of Y and W, and of their *separate* functional PCA decompositions (equivalently, separating the component in X attributable to Y from that attributable to W) on the basis of n curves sampled discretely at a grid of resolution K.

We formulate the problem rigorously in Section 2. Though it might seem that a smooth-plus-rough decomposition is neither unique nor identifiable (except under parametric conditions), we demonstrate in Section 3 that under nonparametric conditions on the covariances of Y and W, such a decomposition is indeed unique (Section 3.1, Theorem 1) and moreover identifiable on the basis of discrete measurements (Section 3.2, Theorem 2). These elucidate the interplay of rank, scale and grid resolution. Estimators of the covariances of Y and W (without assuming knowledge of the rank r and scale δ) are then constructed in Section 4 by means of band deletion and low rank matrix completion using nonlinear least squares (combining smoothing and dimension reduction into a single step). Their asymptotic behaviour is studied in Section 6. These estimates are then used in Section 5 to recover the separate functional PCAs of the Y_i and the W_i, producing a separation of the two scales of variation. The finite sample performance of the methodology is investigated by means of a simulation study in Section 8.

2. Problem statement.

Let $X : [0, 1] \to \mathbb{R}$ be a mean-zero mean square continuous random function, viewed as a random element of the space of integrable real functions defined on $[0, 1]$, say $L^2([0, 1])$, with the usual inner product and induced norm

$$\langle f, g\rangle_{L^2} = \int_0^1 f(t)g(t)\,dt \quad \text{and} \quad \|f\|_{L^2}^2 = \langle f, f\rangle_{L^2}.$$

Assume that X can be decomposed as

$$X(t) = Y(t) + W(t), \qquad t \in [0, 1], \tag{2.1}$$

where Y and W are *uncorrelated* random functions corresponding to a "smooth" and a "rough" component, respectively. This implies an additive decomposition of X's covariance operator $\mathscr{R}$, and of its integral kernel $\rho(s,t) = \mathbb{E}[X(s)X(t)]$, as

$$\mathscr{R} = \mathscr{L} + \mathscr{B}, \tag{2.2}$$

$$\rho(s,t) = \ell(s,t) + b(s,t), \qquad s,t \in [0,1], \tag{2.3}$$

respectively, where the terms on the right are the covariance operators, and kernels, of Y and W, respectively:

$$\ell(s,t) = \mathbb{E}\big[Y(s)Y(t)\big] - \mathbb{E}\big[Y(s)\big]\mathbb{E}\big[Y(t)\big], \tag{2.4}$$

$$b(s,t) = \mathbb{E}\big[W(s)W(t)\big] - \mathbb{E}\big[W(s)\big]\mathbb{E}\big[W(t)\big]. \tag{2.5}$$

We will understand the smoothness in Y to represent smooth variation of X, that is, large scale variation occurring over the entire [0, 1]. On the other hand, the roughness of W corresponds to variations that occur at *scales distinctly smaller* than the global scale [0, 1], but not necessarily the instantaneous time scale that characterizes white noise: variation that is smooth only at *shorter time scales*.

Heuristically, if $\mathscr{B}$ is to capture variation at short time scales only, say at scales of order $\delta \in (0,1)$, we expect its kernel to vanish outside a band of size δ,

$$b(s,t) = 0 \qquad \forall |s-t| \geq \delta.$$

Of course, it will still admit a Mercer decomposition

$$b(s,t) = \sum_{j=1}^{\infty} \beta_j \psi_j(s)\psi_j(t) = \mathbf{1}\big\{|t-s| < \delta\big\} \sum_{j=1}^{\infty} \beta_j \psi_j(s)\psi_j(t),$$

for an orthonormal system of eigenfunctions $\{\psi_j\}$. On the other hand, since $\mathscr{L}$ captures global and smooth variation features, it cannot be allowed to have localised eigenfunctions: these should be smooth enough to be *essentially global*. At the same time, they should be finitely many, otherwise they may still succeed in spanning local variations.[2] We thus postulate that

$$\ell(s,t) = \sum_{j=1}^{r} \lambda_j \eta_j(s)\eta_j(t),$$

for $r < \infty$ and for $\{\eta_j\}_{j=1}^{r}$ sufficiently smooth orthonormal functions in $L^2[0,1]$. We will refer to the operator $\mathscr{L}$ as the *smooth operator*, and to $\mathscr{B}$ as the *banded operator*.

[2] Since there exist infinitely smooth orthonormal systems that are complete in $L^2[0,1]$. To be more precise, what one needs is an exponential rate of decay of the eigenvalues $\{\lambda_j\}$, rather than a precisely finite rank, but we will see in Section 3 that a fast rate of decay alone would not suffice for identifiability to hold.

In summary, our setup is

$$\rho(s,t)=\sum_{j=1}^{r}\lambda_j\eta_j(s)\eta_j(t)+\sum_{j=1}^{\infty}\beta_j\psi_j(s)\psi_j(t),$$

where: (1) $0<\delta<1$; (2) $r<\infty$; (3) the $\{\eta_j\}$ are sufficiently smooth. The statistical problem then is: given K discrete measurements on each of n independent copies of X,

$$X_{ij}=X_i(t_j)=Y_i(t_j)+W_i(t_j),\qquad i=1,\dots,n,$$

obtained by point evaluation at some grid points $\{t_1,\dots,t_K\}$:

1. estimate the components $\mathscr{L}$ and $\mathscr{B}$, and their spectral decomposition, and
2. construct separate functional PCAs for the smooth and rough components $\{Y_i\}_{i=1}^n$ and $\{W_i\}_{i=1}^n$ on the basis of these estimates (effectively separating the two scales of variation and recovering the Y_i and W_i).

To do so, we will need to formulate more precise conditions on the smoothness and roughness of the two components, or equivalently the rank and scale of these variations, as it is clear that the problem can otherwise be severely ill-posed (in a sense, the problem can be seen as an infinite-dimensional version of *density estimation with contamination by measurement error of an unknown distribution*, also known as *double-blind deconvolution*). This is done next, in Section 3.

3. Well-posedness: Uniqueness and identifiability.

3.1. *Uniqueness of the decomposition $\mathscr{R}=\mathscr{L}+\mathscr{B}$.* An obvious challenge with a decomposition of the form $\mathscr{R}=\mathscr{L}+\mathscr{B}$, is that there may be infinitely many distinct pairs $(\mathscr{L},\mathscr{B})$ whose sum yields the same $\mathscr{R}$: we are asking to identify two summands from knowledge of their sum. As it turns out, uniqueness is a matter of scale: assuming that variations of the W process propagate only locally, at most at scale δ, whereas that variations of Y are purely nonlocal. The next theorem makes this statement precise via the notion of *analyticity*.

THEOREM 1 (Uniqueness). *Let $\mathscr{L}_1,\mathscr{L}_2:L^2[0,1]\to L^2[0,1]$ be trace-class covariance operators of rank $r_1<\infty$ and $r_2<\infty$, respectively. Let $\mathscr{B}_1,\mathscr{B}_2:L^2[0,1]\to L^2[0,1]$ be banded trace-class covariance operators of bandwidth $\delta_1<1$ and $\delta_2<1$, respectively. If the eigenfunctions of $\mathscr{L}_1$ and $\mathscr{L}_2$ are real analytic, then we have the equivalence*:

$$\mathscr{L}_1+\mathscr{B}_1=\mathscr{L}_2+\mathscr{B}_2\quad\Longleftrightarrow\quad\mathscr{L}_1=\mathscr{L}_2\quad\textit{and}\quad\mathscr{B}_1=\mathscr{B}_2.$$

REMARK 1 (Sufficiency vs. necessity). The conditions of the theorem can actually be strictly weakened, with the same conclusion: instead of requiring finite ranks and analytic eigenfunctions for $(\mathscr{L}_1,\mathscr{L}_2)$, it suffices to require the weaker condition that their kernels be analytic on an open set $U \subset [0,1]^2$ that contains the larger of the two bands, $U \supset \{(s,t) \in [0,1]^2 : |t-s| \leq \max(\delta_1, \delta_2)\}$. This can be relaxed no further, though: if the kernels of $(\mathscr{L}_1, \mathscr{L}_2)$ are not analytic on such a U, one can construct counterexamples, at least at this level of generality. For such counterexamples, see the Supplementary Material [8], Section 3. Thus analyticity is necessary, unless further assumptions are imposed on the banded covariances. We choose to put the spotlight on the stronger assumption of the finite rank analytic eigenfunction case, because: (a) this is the one that will be practically relevant in light of the identifiability conditions that will be established in Section 3.2 (Theorem 2), and (b) the set of rank r covariance operators with analytic eigenfunctions is a dense subset of the set of all rank r covariance operators (see Proposition 1 below), giving us a rich set of identifiable models of the form (2.1).

Recall that a function is real analytic on an open interval if and only if its Fourier coefficients decay at a rate that is at least geometric (see Krantz and Parks [13] for a detailed survey of real analytic functions). For instance, if we write $\eta(x) = \sum_{k=1}^{\infty} (\alpha_k \cos(kx) + b_k \sin(kx))$, then η is real analytic on $(-\pi, \pi)$ if an only if

$$\limsup_{k \to \infty} \bigl(|\alpha_k| + |\beta_k|\bigr)^{1/k} < 1.$$

Examples of analytic functions include polynomials, trigonometric functions, exponential and logarithmic functions, rational functions with no poles, truncated Gaussians and finite location/scale mixtures thereof, to name only a few; such functions have been routinely used as typical examples of low order eigenfunctions capturing smooth variation in functional data analysis. The class of real analytic functions is also closed under finite linear combination, multiplication and division (assuming a nonvanishing denominator), composition, differentiation and integration. Thus, one can generate rich collections of analytic eigenfunctions (and hence analytic covariance operators) by combining analytic functions. In fact, the set of rank r covariance operators with analytic eigenfunctions is a dense subset of the set of all rank r covariance operators:

PROPOSITION 1. *Let Z be an $L^2[0,1]$-valued random function with a trace class covariance $\mathscr{G}$ of rank $r < \infty$. Then, for any $\varepsilon > 0$ there exists a random function Y whose covariance $\mathscr{L}$ has analytic eigenfunctions and rank $q \leq r$, such that*

$$\mathbb{E}\|Z - Y\|_{L^2}^2 < \varepsilon \quad \textit{and} \quad \|\mathscr{G} - \mathscr{L}\|_* < \varepsilon,$$

for $\|\cdot\|_*$ the nuclear norm. If additionally $\mathscr{G}$ has C^1 eigenfunctions on $[0,1]$, then we have the stronger result that for any $\varepsilon > 0$, there exists a random function Y whose covariance $\mathscr{L}$ has analytic eigenfunctions and rank $q \leq r$, such that

$$\sup_{t \in [0,1]} \mathbb{E}\bigl|Z(t) - Y(t)\bigr|^2 < \varepsilon \quad \textit{and} \quad \sup_{s,t \in [0,1]} \bigl|g(s,t) - \ell(s,t)\bigr| < \varepsilon,$$

where g and ℓ are the kernels of $\mathscr{G}$ and $\mathscr{L}$, respectively.

Note that an immediate conclusion is that, for a given r, the accuracy of a rank r analytic approximation of a mean-square continuous process can be made arbitrarily close to the accuracy of the (optimal) rank r Karhunen–Loève approximation, in the same uniform mean square sense. Thus, if we expect a process to be approximately of low rank r (as in our model of Section 2), then this process can be very well approximated by an analytic process of the same low rank r. This shows that the condition of analyticity, at least as a model that guarantees uniqueness of decomposition $\mathscr{R} = \mathscr{L} + \mathscr{B}$, is not nearly as restrictive as it may seem at first sight (and in any case, it is sharp given the discussion in Remark 1).

3.2. *Identifiability at finite resolution.* Theorem 1 relies on an analyticity assumption, which is a fundamentally functional assumption, so it is not clear whether the result is useful in practice: is the decomposition identifiable on the basis of finitely many discrete measurements? Remarkably the answer is yes, and crucially depends both on the finite rank and the analyticity assumption.

Suppose we are given K discrete measurements on each of n independent copies of X,

$$X_{ij} = X_i(t_j) = Y_i(t_j) + W_i(t_j), \qquad i = 1, \dots, n,$$

obtained by evaluation at points $\{t_j\}_{j=1}^K$, where

$$(t_1, \dots, t_K) \in \mathcal{T}_K = \{(x_1, \dots, x_K) \in \mathbb{R}^K : x_1 \in I_{1,K}, \dots, x_K \in I_{K,K}\},$$

and $\{I_{j,K}\}_{j=1}^K$ is the partition of [0, 1] into intervals of length $1/K$. With this information, we can of course only hope to be able to uniquely identify the K-resolution versions of the operators, $(\mathscr{L}, \mathscr{B})$, say $(\mathscr{L}^K, \mathscr{B}^K)$ on the basis of the K-resolution version of their sum, say $\mathscr{R}^K = \mathscr{L}^K + \mathscr{B}^K$. These operators are defined to have kernels

$$\rho^K(x, y) = \sum_{i,j=1}^K \rho(t_i, t_j)\mathbf{1}\{(x, y) \in I_{i,K} \times I_{j,K}\}, \tag{3.1}$$

$$\ell^K(x, y) = \sum_{i,j=1}^K \ell(t_i, t_j)\mathbf{1}\{(x, y) \in I_{i,K} \times I_{j,K}\}, \tag{3.2}$$

$$b^K(x, y) = \sum_{i,j=1}^K b(t_i, t_j)\mathbf{1}\{(x, y) \in I_{i,K} \times I_{j,K}\}, \tag{3.3}$$

which can be summarised via the following $K \times K$ matrix representations:

$$R^K(i, j) = \rho(t_i, t_j), \qquad L^K(i, j) = \ell(t_i, t_j), \qquad B^K(i, j) = b(t_i, t_j).$$

Without loss of generality, one can assume that R^K has been re-normalised to be of unit trace norm, whenever convenient. As it turns out, there exists a finite critical

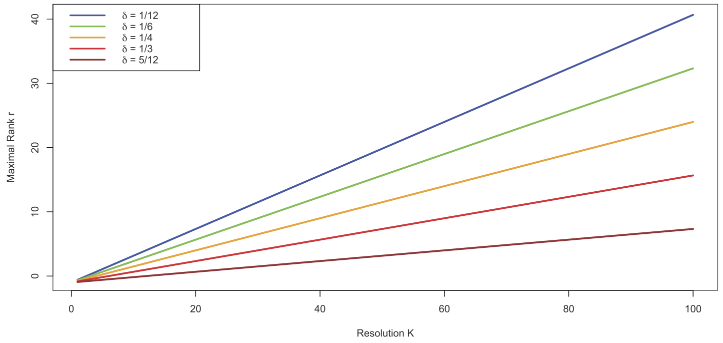

FIG. 1. *Graphic representation of the interplay between rank, scale and resolution. For different values of the scale parameter δ, the maximal identifiable rank r is plotted as a function of the resolution K.*

resolution K^*, with explicit dependence on the rank r and scale δ, beyond which identification is possible, provided that $r < \infty$ and $\delta < 1/2$. This encapsulates the interplay between rank, resolution and scale.

THEOREM 2 (Discrete identifiability). *Let $\mathscr{L}_1$ and $\mathscr{L}_2$ be covariance operators of finite ranks $r_1 < \infty$ and $r_2 < \infty$, respectively, and assume without loss of generality that $r_1 \geq r_2$. Let $\mathscr{B}_1$ and $\mathscr{B}_1$ be two banded continuous covariance operators of bandwidth $\delta_1 < 1/2$ and $\delta_2 < 1/2$, respectively. Given $(t_1, \ldots, t_K) \in \mathcal{T}_K$, define their K-resolution matrix coefficients to be $(L_1^K, B_1^K, L_2^K, B_2^K) \in \mathbb{R}^{K \times K}$,*

$$L_m^K(i, j) = \ell_m(t_i, t_j) \quad \textit{and} \quad B_m^K(i, j) = b_m(t_i, t_j), \qquad i, j \in \{1, \ldots, K\},$$

for $m = 1, 2$. If the eigenfunctions of $\mathscr{L}_1$ and $\mathscr{L}_2$ are all real analytic, and

$$K \geq K^* = \max\left(\frac{2r_1 + 2}{1 - 2\delta_1}, \frac{2r_1 + 2}{1 - 2\delta_2}\right),$$

then we have the equivalence

$$L_1^K + B_2^K = L_2^K + B_2^K \quad \Longleftrightarrow \quad L_1^K = L_2^K \quad \textit{and} \quad B_1^K = B_2^K,$$

almost everywhere on $\mathcal{T}_K$ with respect to Lebesgue measure.

The theorem reveals the interplay between the fundamental parameters of the problem, which is governed by the constraint:

$$r \leq \left(\frac{1}{2} - \delta\right)K - 1. \tag{3.4}$$

This yields the maximal rank that the smooth operator can have, for a given resolution K and scale δ of the banded operator, if the problem is to be identifiable. Figure 1 plots this maximal rank r as a function of K for different values of the parameter δ. We note that things are not particularly restrictive, allowing identifiability for quite large values of the bandwidth δ and rather modest values of K, when the rank r is not exceedingly large, as is nearly always assumed in the practice of FDA.

An attractive feature of this result is that the conditions imposed are deterministic and yet not particularly restrictive. This is in contrast with results in recent progress on matrix completion which either have restrictive deterministic conditions, or more relaxed but random conditions. The reason is that we are fortunate to have a deterministic and known structure of the missing set of values to be completed.

The main caveat of passing from the continuum to discrete observation is that the theorem is valid almost everywhere on $\mathcal{T}_K$, rather than pointwise on $\mathcal{T}_K$. Thus, we know that the identifiability holds for *almost all grids* without being able to conclusively say so for *a specific grid*. In probabilistic terms, if the points t_j are chosen independently at random, each according to an absolutely continuous distribution on the corresponding interval I_j, then we know that identifiability holds with probability 1.

4. Estimation by matrix completion.

Our strategy for estimation will be to define an objective function depending only on R^K whose unique optimum yields the required matrix L^K. Then we will define an estimator of L^K on the basis of an empirical version of this objective function. Ideally, the objective function should not depend on the knowledge of the unknown quantities δ and r, otherwise there would be two "competing" tuning parameters to choose. The following proposition yields such an objective function, in the form of a low rank matrix completion problem.

PROPOSITION 2. *Let $\mathscr{L} : L^2[0,1] \to L^2[0,1]$ be a rank $r < \infty$ covariance operator with analytic eigenfunctions and kernel ℓ, and $\mathscr{B} : L^2[0,1] \to L^2[0,1]$ a trace-class covariance operator with δ-banded kernel b. For $(t_1,\dots,t_K) \in \mathcal{T}_K$, let*

$$L^K = \{\ell(t_i,t_j)\}_{ij}, \qquad B^K = \{b(t_i,t_j)\}_{ij},$$

and $R^K = L^K + B^K$. Assume that

$$\delta < \frac{1}{4} \quad \textit{and} \quad K \geq 4r + 4.$$

Define the matrix $P^K \in \mathbb{R}^{K\times K}$ by $P^K(i,j) = \mathbf{1}\{|i-j| > \lceil K/4 \rceil\}$. Then, for almost all grids in $\mathcal{T}_K$:

1. *The matrix L^K is the unique solution to the optimization problem*

$$\min_{\theta\in\mathbb{R}^{K\times K}} \operatorname{rank}\{\theta\} \quad \textit{subject to} \quad \|P^K \circ (R^K - \theta)\|_F^2 = 0. \tag{4.1}$$

2. *Equivalently, in penalised form,*

$$L^K = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{K\times K}} \{ \| P^K \circ (R^K - \theta) \|_F^2 + \tau \operatorname{rank}(\theta) \}, \tag{4.2}$$

for all $\tau > 0$ sufficiently small.

Here, $\|\cdot\|_F$ is the Frobenius matrix norm and "$\circ$" denotes the Hadamard product.

Simply put, among all possible matrix completions of $P^K \circ (R^K - \theta)$, the matrix L^K is uniquely the one of lowest rank: no matrix of rank lower than the true rank r will provide a completion; and any completion other than L^K will have rank at least $r+1$. Notice that neither of the objective functions (4.1) or (4.2) depends on δ or r: unique recovery of L^K and B^K is feasible even when we do not know the true values of r or δ. The concession we had to make to achieve this adaptation is to require $\delta < 1/4$ (compared to $\delta < 1/2$ in Theorem 2). In particular, we use the penalised form in equation (4.2) to motivate the formal definition of our estimation approach [the equivalent form in equation (4.1) will be useful for computation, see Section 7]:

DEFINITION 1 (Estimator of L^K). Let $(X_1, \dots, X_n)$ be i.i.d. copies of $X = Y + W$. Let $(t_1, \dots, t_K) \in \mathcal{T}_K$ and assume we observe

$$X_{ij} = X_i(t_j), \qquad i = 1, \dots, n; j = 1, \dots, K.$$

Let $R_n^K \in \mathbb{R}^{K\times K}$ be the empirical covariance matrix of the vectors

$$\{(X_{i1}, \dots, X_{iK})^\top\}_{i=1}^n.$$

We define the estimator $\hat{L}_n^K$ of L^K to be an approximate minimum of

$$\min_{\theta \in \Theta_K} \{ K^{-2} \| P^K \circ (R_n^K - \theta) \|_F^2 + \tau \operatorname{rank}(\theta) \}, \tag{4.3}$$

where $P^K \in \mathbb{R}^{K\times K}$ is defined as $P^K(i,j) = \mathbf{1}\{|i-j| > \lceil K/4 \rceil\}$, $\tau > 0$ is a sufficiently small tuning parameter, and Θ_K is the set of $K \times K$ nonnegative matrices of trace norm bounded by that of R_n^K (which can be renormalised to unit trace norm). By approximate minimum, it is meant that the value of the functional at $\hat{L}_n^K$ is within $O_{\mathbb{P}}(n^{-1})$ of the value of the overall minimum.

We discuss the practical implementation of the estimation method of Definition 1, including the selection of the tuning parameter, in Section 7. Once $\hat{L}_n^K$ has been constructed, we may also construct a plug-in estimator for B^K.

DEFINITION 2 (Plug-in estimator of B^K). Let R_n^K and $\hat{L}_n^K$ be as in Definition 1. We define the plug-in estimator $\hat{B}_n^K$ of B_n^K to be the projection of $\Delta_n^K = R_n^K - \hat{L}_n^K$ onto the convex set of nonnegative banded $K \times K$ matrices of bandwidth at most $\lceil K/4 \rceil$.

We could of course have used $\Delta_n^K = R_n^K - \hat{L}_n^K$ itself to estimate B^K, but there is no guarantee that this will be positive definite. Asymptotically in n, Δ_n^K and $\hat{B}_n^K$ will coincide. Note that the intersection of the set of banded matrices (with given band) and the set of nonnegative matrices is a closed convex set, thus the projection uniquely exists. In practice, it can be approximately determined by the method of alternative projections, or Dykstra's algorithm (see Section 7).

Once $\hat{L}_n^K$ and $\hat{B}_n^K$ are at hand, it is reasonable to use their sum as an estimator of R^K, instead of the empirical version R_n^K, as the former is in principle less "noisy" than the latter.

DEFINITION 3 (Plug-in estimator of R^K). Let $\hat{L}_n^K$ and $\hat{B}_n^K$ be as in Definitions 1 and 2. We define the plug-in estimator $\hat{R}_n^K$ of R^K as $\hat{R}_n^K = \hat{L}_n^K + \hat{B}_n^K$.

Our K-resolution estimators $(\hat{\mathscr{L}}_n^K, \hat{\mathscr{B}}_n^K, \hat{\mathscr{R}}_n^K)$ of $(\mathscr{L}, \mathscr{B}, \mathscr{R})$ will now be defined as the operators with step-function kernels $[\hat{\ell}_n^K(x, y), \hat{b}_n^K(x, y), \hat{\rho}_n^K(x, y)]$ whose coefficients are given by the matrices $(\hat{L}_n^K, \hat{B}_n^K, \hat{R}_n^K)$:

$$\hat{\ell}_n^K(x, y) = \sum_{j=1}^K \hat{L}_n^K(i, j)\mathbf{1}\{(x, y) \in I_{i,K} \times I_{j,K}\},$$

$$\hat{b}_n^K(x, y) = \sum_{j=1}^K \hat{B}_n^K(i, j)\mathbf{1}\{(x, y) \in I_{i,K} \times I_{j,K}\},$$

$$\hat{\rho}_n^K(x, y) = \sum_{j=1}^K \hat{R}_n^K(i, j)\mathbf{1}\{(x, y) \in I_{i,K} \times I_{j,K}\}.$$

Correspondingly, the estimators of their spectra will be given by the spectra of $\hat{\mathscr{L}}_n^K$, $\hat{\mathscr{B}}_n^K$, and $\hat{\mathscr{R}}_n^K$:

$$\hat{\mathscr{L}}_n^K = \sum_{j=1}^{\hat{r}} \hat{\lambda}_j \hat{\eta}_j \otimes \hat{\eta}_j, \qquad \hat{\mathscr{B}}_n^K = \sum_{j=1}^{K} \hat{\beta}_j \hat{\psi}_j \otimes \hat{\psi}_j, \qquad \hat{\mathscr{R}}_n^K = \sum_{j=1}^{K} \hat{\theta}_j \hat{\varphi}_j \otimes \hat{\varphi}_j.$$

Here, $\hat{r} \leq K/4$ is the rank of $\hat{\mathscr{L}}_n^K$. Note that the empirical eigenfunctions $\hat{\eta}_j$ of $\hat{\mathscr{L}}_n^K$ will be step functions. They can, of course, be replaced by smooth versions thereof. For example, one can smooth the covariance function $\hat{\ell}_n^K$, and then calculate the spectrum of the induced covariance operator. The amount of smoothing required will be rather limited since $\hat{\ell}_n^K$ is effectively already de-noised. One could also directly smooth the eigenfunctions, but then there is no guarantee that their smoothed versions will be still orthogonal. Without any additional smoothness assumptions on $\mathscr{B}$, we cannot presume to smooth the step functions $\hat{\psi}_j$ in order to obtain smoother versions (recall that only continuity of b was assumed).

5. Separation of scales. With estimators of the covariance operators $(\mathscr{L}, \mathscr{B})$ and their spectra at our disposal, we now wish to carry out functional PCA separately for the smooth and the rough components, thus separating the two scales of variation. In order to have identifiability at the level of curves, we need to add the assumption that at least one of the two processes Y and W has a known mean. Here, we assume that the rough process W is known to have mean zero, and to simplify the presentation we assume that the mean of Y has been removed from the data so we have $\mathbb{E}[Y] = 0$, too. Focussing on the smooth component, we note that its Karhunen–Loève expansion is

$$Y_i = \sum_{j=1}^{r} \langle Y_i, \eta_j \rangle \eta_j.$$

Having estimated η_j already, it suffices to estimate the scores $\{\langle Y_i, \eta_j \rangle\}_{i=1}^{n}$, in order to have a complete analysis into principal components. If we were able to observe $\{Y_i(t_j)\}_{i,j}$, then the natural estimator would be given by

$$\langle Y_i^K, \hat{\eta}_j \rangle_{L^2} = \frac{1}{K} \sum_{k=1}^{K} Y_i(t_k) \hat{\eta}_j(t_k),$$

where $Y_i^K(t) = \sum_{j=1}^{K} Y_i(t_j) \mathbf{1}\{t \in I_{j,K}\}$. A parallel discussion holds in the case of the rough components $\{W_i\}$. In effect, we see that the problem of estimating the principal scores of Y and W separately is equivalent to that of *separating* the unobservable components $Y_i(t_j)$ and $W_i(t_j)$ in the decomposition

$$X_i(t_j) = Y_i(t_j) + W_i(t_j),$$

on the basis of the observations $X_i(t_j)$. We concentrate on a specific observation, say $i = 1$, and drop the index 1 for the sake of tidiness.

Separation can be viewed as a problem of *prediction* (similar to the approach taken by Yao et al. [21]). If the covariance operators $\mathscr{R}$ and $\mathscr{L}$ were known precisely, then we would attempt to recover the components $Y^K(t) = \sum_{j=1}^{K} Y(t_j) \mathbf{1}\{t \in I_{j,K}\}$ and $W^K(t) = \sum_{j=1}^{K} W(t_j) \mathbf{1}\{t \in I_{j,K}\}$ by means of their best predictors given the observation $X^K(t) = \sum_{j=1}^{K} X(t_j) \mathbf{1}\{t \in I_{j,K}\}$. The most tractable case is that of using the best *linear* predictor (which is best overall in the Gaussian case), and this is what we will pursue. Noting that Y and W are zero mean and uncorrelated, the best linear predictor of Y^K given X^K (viewed as random elements of L^2) is

$$\Pi(X^K) = \sum_{j=1}^{r} \sum_{i=1}^{q} \frac{\lambda_j^K}{\theta_i^K} \langle \varphi_i^K, \eta_j^K \rangle \langle \varphi_i^K, X^K \rangle \eta_j^K = \sum_{j=1}^{r} \xi_j \eta_j^K, \tag{5.1}$$

where $\{\theta_i^K, \varphi_i^K\}_{i=1}^{q}$ is the spectrum of $\mathscr{R}^K$ (with $q \leq \infty$) and $\{\lambda_j^K, \eta_j^K\}_{j=1}^{r}$ that of $\mathscr{L}^K$ (see Bosq [3], Proposition 3.1, and Bosq [3], Example 3.3). Note that $\mathscr{R}^K$ is the covariance operator of X^K.

We estimate the best linear predictor, by replacing the unknown elements in equation (5.1) by their corresponding estimators. Specifically, recalling that

$$\hat{\mathscr{R}}_n^K = \sum_{i=1}^{\hat{q}} \hat{\theta}_i \hat{\varphi}_i \otimes \hat{\varphi}_i, \qquad \hat{q} = \operatorname{rank}(\hat{\mathscr{R}}_n^K) \quad \text{and}$$

$$\hat{\mathscr{L}}_n^K = \sum_{j=1}^{\hat{r}} \hat{\lambda}_j \hat{\eta}_j \otimes \hat{\eta}_j, \qquad \hat{r} = \operatorname{rank}(\hat{\mathscr{L}}_n^K),$$

our estimator of the predictor of Y^K given X^K is

$$\hat{Y}_n^K := \sum_{j=1}^{\hat{r}} \sum_{i=1}^{\hat{q}} \frac{\hat{\lambda}_j}{\hat{\theta}_i} \langle \hat{\varphi}_i, \hat{\eta}_j \rangle \langle \hat{\varphi}_i, X^K \rangle \hat{\eta}_j = \sum_{j=1}^{\hat{r}} \hat{\xi}_j \hat{\eta}_j. \tag{5.2}$$

In matrix notation, the estimated scores $(\hat{\xi}_1, \dots, \hat{\xi}_{\hat{r}})^\top$ of Y satisfy

$$\hat{\xi}_j = \langle \hat{\lambda}_j (\hat{\mathscr{R}}_n^K)^\dagger \hat{\eta}_j, X^K \rangle = \frac{1}{K} \hat{\lambda}_j \mathbf{X}^\top (\hat{R}_n^K)^\dagger \hat{\boldsymbol{\eta}}_j = \frac{1}{K} \hat{\lambda}_j \mathbf{X}^\top (\hat{L}_n^K + \hat{B}_n^K)^\dagger \hat{\boldsymbol{\eta}}_j, \tag{5.3}$$

where $\mathbf{X} = (X(t_1), \dots, X(t_K))^\top$, $\hat{\boldsymbol{\eta}}_j = (\hat{\eta}_j(t_1), \dots, \hat{\eta}_j(t_K))^\top$, and we use the notation $\mathscr{A}^\dagger$ to denote the generalised inverse of an operator (or matrix) $\mathscr{A}$. It is worth remarking that the last expression in equation (5.3) is essentially the same as that of the PACE estimator of Yao et al. [21], with the exception that one has a banded matrix $\hat{B}_n^K$ in lieu of a diagonal matrix of the form $\hat{\sigma}^2 I$. The best linear predictor of W^K given X^K, say $\Psi(X^K)$, can be estimated by means of the *residuals*

$$\hat{W}(t_j) = X(t_j) - \hat{Y}_n^K(t_j), \qquad j = 1, \dots, K.$$

This definition is motivated from the simple fact that

$$\Psi(X^K) = \mathbb{E}[W^K | X^K] = \mathbb{E}[X^K - Y^K | X^K] = X^K - \mathbb{E}[Y^K | X^K] = X^K - \Pi(X^K).$$

6. Asymptotic theory.

We now turn to consider the asymptotic behaviour of the estimators constructed in the last two sections. Our first result considers the asymptotic behaviour of our estimator $\hat{\mathscr{L}}_n^K$ and its spectrum, in terms of the observation grid and the number of curves. In the sequel, we will follow the usual convention that the sign of the estimated eigenfunctions is correctly identified (since only the eigenprojectors are formally identifiable).

THEOREM 3. *In the setting of Section 4, let the $r < \infty$ eigenvalues of $\mathscr{L}$ be of multiplicity one, $\mathbb{E}\|X\|_{L^2}^4 < \infty$ and $\delta < \frac{1}{4}$, and define $K^* = 4(r+1)$ to be the critical resolution. Then, for any $K > K^*$ and almost all grids in $\mathcal{T}_K$ it holds that*

$$\|\hat{\mathscr{L}}_n^K - \mathscr{L}\|_{\mathrm{HS}}^2 \le O_{\mathbb{P}}(n^{-1}) + 4K^{-2} \sup_{x,y \in [0,1]} \|\nabla \ell(x, y)\|_2^2, \tag{6.1}$$

$$\|\hat{\eta}_j - \eta_j\|_{L^2}^2 \leq O_{\mathbb{P}}(n^{-1}) + 2K^{-2}\|\eta_j'\|_\infty^2, \qquad j \in \{1, \dots, r\}, \tag{6.2}$$

$$\sup_{j\geq 1} |\hat{\lambda}_j - \lambda_j|^2 \leq O_{\mathbb{P}}(n^{-1}) + 4K^{-2} \sup_{x,y\in[0,1]} \|\nabla \ell(x, y)\|_2^2, \tag{6.3}$$

for all $\tau > 0$ *sufficiently small, where* $\|\cdot\|_{\mathrm{HS}}$ *is the Hilbert–Schmidt norm of an operator. Furthermore, the rank of* $\hat{\mathscr{L}}_n^K$ *satisfies*

$$\left|\operatorname{rank}(\hat{\mathscr{L}}_n^K) - r\right| = O_{\mathbb{P}}(n^{-1}). \tag{6.4}$$

REMARK 2. The fact that the theorem holds true almost everywhere on $\mathcal{T}_K$ can equivalently be stated in probabilistic terms. Assume that the grid $\mathbf{t}_K = \{t_{j,K}\}_{j=1}^K$ is chosen at random according to the uniform distribution on $\mathcal{T}_K$. Then the theorem holds with probability 1 over the grid choice. Note that the uniform measure on $\mathcal{T}_K$ can be generated by selecting $\{t_{j,K}\}_{j=1}^K$ to be independent for $j \in \{1, \dots, K\}$, each uniformly distributed on the corresponding subinterval $I_{j,K}$.

Similar asymptotics for $\hat{\mathscr{B}}_n^K$ follow as a corollary, since it is defined as a contraction of the difference $\mathscr{R}_n^K - \hat{\mathscr{L}}_n^K$.

COROLLARY 1. *Assume that the eigenvalues of* $\mathscr{B}$ *are of multiplicity one. If the covariance function* $b(s,t) : [0, 1]^2 \to \mathbb{R}$ *associated with* $\mathscr{B}$ *is continuously differentiable, then under the same conditions as in Theorem* 3, *and for any* $K > K^*$ *and almost all grids in* $\mathcal{T}_K$ *we have*

$$\|\hat{\mathscr{B}}_n^K - \mathscr{B}\|_{\mathrm{HS}}^2 \leq O_{\mathbb{P}}(n^{-1}) + 4K^{-2} \sup_{x,y\in[0,1]} \|\nabla b(x, y)\|_2^2, \tag{6.5}$$

$$\frac{\sigma_j^2}{8}\|\hat{\psi}_j - \psi_j\|_{L^2}^2 \leq O_{\mathbb{P}}(n^{-1}) + \frac{\sigma_j^2}{4}K^{-2}\|\psi_j'\|_\infty^2, \tag{6.6}$$

$$\sup_{j\geq 1} |\hat{\beta}_j - \beta_j|^2 \leq O_{\mathbb{P}}(n^{-1}) + 4K^{-2} \sup_{x,y\in[0,1]} \|\nabla b(x, y)\|_2^2, \tag{6.7}$$

for all $\tau > 0$ *sufficiently small. Here*

$$\sigma_1 = \beta_1 - \beta_2 \quad \textit{and} \quad \sigma_j = \min\{\beta_{j-1} - \beta_j, \beta_j - \beta_{j+1}\}, \qquad 2 \leq j \leq \operatorname{rank}(\mathscr{B}) \wedge K,$$

The last two results can now be combined to obtain the asymptotic behaviour of $\hat{\mathscr{R}}_n^K$.

COROLLARY 2. *Under the same conditions as in Theorem* 3 *and Corollary* 1, *we have that for any* $K > K^*$ *and almost all grids in* $\mathcal{T}_K$,

$$\|\hat{\mathscr{R}}_n^K - \mathscr{R}\|_{\mathrm{HS}}^2 \leq O_{\mathbb{P}}(n^{-1}) + 4K^{-2} \sup_{x,y\in[0,1]} \|\nabla \rho(x, y)\|_2^2, \tag{6.8}$$

for all $\tau > 0$ *sufficiently small.*

Finally, we show that the predictors of Y^K and W^K based on a finite grid of resolution K are consistent in the L^2 sense, which also implies that the corresponding estimated PCA scores are consistent, too.

COROLLARY 3. *In the same setting as in Theorem 3, let $K > K^*$. If $\mathscr{R}^K$, is of full rank, and if the kernel $b(s,t) : [0,1]^2 \to \mathbb{R}$ of $\mathscr{B}$ is continuously differentiable, then*

$$\|\hat{Y}_n^K - \Pi(X^K)\|_{L^2} = O_{\mathbb{P}}(n^{-1/2}),$$
$$\|\hat{W}_n^K - \Psi(X^K)\|_{L^2} = O_{\mathbb{P}}(n^{-1/2}),$$

almost everywhere on $\mathcal{T}_K$.

7. Practical implementation via band-deleted PCA. To compute the estimators $\hat{L}_n^K$ and $\hat{B}_n^K$ from a sample of discretely observed curves $\mathbf{X_1}, \dots, \mathbf{X}_n$, where $\mathbf{X}_i = (X_i(t_1), \dots, X_i(t_K))^\top$, we apply the following algorithm:

(A) Compute the empirical covariance matrix of the sample

$$R_n^K = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \hat{\boldsymbol{\mu}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^\top \qquad \text{where } \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i.$$

(B) Solve the optimisation problem

$$\min_{0 \preceq \theta \in \mathbb{R}^{K \times K}} \|P^K \circ (R_n^K - \theta)\|_F^2 \quad \text{subject to} \quad \operatorname{rank}(\theta) \le i, \tag{7.1}$$

for $i = \{1, \dots, K/4 - 1\}$, obtaining minimisers $\hat{\theta}_1, \dots, \hat{\theta}_{K/4-1}$.

(C) Calculate the *fits* $\{f(i) = \|P^K \circ (R_n^K - \hat{\theta}_i)\|_F^2 : i = 1, \dots, K/4 - 1\}$, and the quantities

$$f(i) + \tau i,$$

for some choice of the tuning parameter $\tau > 0$.

(D) Determine the i that minimises the above quantity, and declare the corresponding optimising matrix to be the estimator $\hat{L}_n^K$.

(E) Use an alternating projection algorithm (Bauschke and Borwein [1]) to compute an approximation of the projection of $R_n^K - \hat{L}_n^K$ onto the intersection of the set of banded $K \times K$ matrices of bandwidth at most $\lceil K/4 \rceil$ and the set of nonnegative definite $K \times K$ matrices. Set the resulting matrix to be $\hat{B}_n^K$.

Notice that τ being positive in step (C) precludes us from overfitting by choosing a matrix of arbitrarily large rank. A natural question is: *how does one choose the precise τ in Step (C)*? The answer is that, any choice of τ implies a choice of rank i_τ (this being the rank of the optimum corresponding to τ), and thus a fit value $f(i_\tau)$. Thus one can use the the *scree-plot $i \mapsto f(i)$* as a guide to implicitly choose τ, by replacing step (C) with:

(C$'$) Plot the nonincreasing function $i \mapsto f(i)$, and choose a value of i to be the smallest one such that $f(i) < c$, for some threshold value c. Then declare the corresponding optimising matrix to be the estimator $\hat{L}_n^K$. Again, c being positive precludes us from overfitting by choosing an arbitrarily large rank.

REMARK 3. The solution of (C$'$) for a certain choice of $c > 0$ is equivalent to the solution of (C) for a certain corresponding choice of τ (when the scree plot has a convex shape, as has been the case in all the simulations we carried out, there is an explicit relationship between c and τ; see the Supplementary Material [8], Section 4).

The value c is in principle chosen to be small (converging to zero as n increases), and corresponds to selecting a value i for the rank beyond which the function f levels out. This is precisely an "elbow selection rule" as is usual with scree-plots in PCA. The analogy with traditional scree plots and PCA is, in fact, quite strong: in traditional PCA, for each i one determines a rank i matrix that best fits the empirical covariance, and then chooses an appropriate i via a scree plot. Here, we do *almost that*: for each i, we determine a rank i matrix that best fits the band-deleted empirical covariance, and then we choose an appropriate i via a scree plot. Particularly in our case, a clear motivation for the "elbow" approach comes from the fact that if we could solve (7.1) with R^K instead of R_n^K, then we would have

$$f(i) > 0 \qquad \text{if } i = 1, \dots, r-1 \quad \text{and} \quad f(i) = 0 \qquad \text{if } i \geq r.$$

The asymptotic validity of this motivation is shown in the Supplementary Material [8], Section 4.

Going back to Step (B), another difference with traditional PCA, is that the best rank i approximation of the off-band elements of the empirical covariance cannot be determined in closed form by simple eigenanalysis. Thus, we must use approximate schemes in order to solve the optimisation problem (7.1). For a given value of i, we use the fact that any $K \times K$ positive semi-definite matrix of rank at most i can be factorised as $CC^\top$, with $C \in \mathbb{R}^{K \times i}$. The problem thus reduces to

$$\min_{C \in \mathbb{R}^{K \times i}} \left\| P^K \circ \left(R_n^K - CC^\top\right) \right\|_F^2, \tag{7.2}$$

for $i = 1, \dots, K/4 - 1$. Notice that these problems are *not* convex in C, and we thus do not have guarantees that gradient descent-type algorithms will converge to a global optimum (of which there are multiple, since the matrix factorisation is not unique). That being said, recent theoretical progress (e.g., Chen and Wainwright [4]) shows that, remarkably, projected gradient descent methods with a reasonable starting point have high probability of yielding "good" local optima in factorised matrix completion problems. In our own implementations for example, in our simulations in Section 8, we solve the optimisation problem (7.2) (which can be seen as factorised matrix completion) using the function `fminunc` of the optimization toolbox in MATLAB [15], with starting point $C_0 = U_i \Sigma_i^{1/2}$, where: $U\Sigma U^T$ is the singular value decomposition of R_n^K; U_i is the $n \times i$ matrix obtained by keeping the first i columns of U; and Σ_i is the $i \times i$ matrix obtained by keeping the first i lines and columns of Σ. This function uses a subspace trust-region method based on the interior-reflective Newton method described in [6] and [5] to perform the optimization. Though we do not use the exact same method, we are in a similar setup as Chen and Wainwright [4], so we can expect to obtain "good" local optima. Indeed, in our simulations (Section 8), the computational method was stable and quickly converged to a reasonable local optimum.

8. Simulation study. In order to study the performance of our method on a broad range of setups, we consider nine general scenarios to simulate our data. For each of these scenarios, we simulate n i.i.d. mean-zero functions Y_i and n i.i.d. mean-zero functions W_i on a grid of K equally spaced points on the interval [0, 1]. From these samples of discretised curves, we calculate the matrices L_n^K and B_n^K:

$$L_n^K(a,b) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t_a)Y_i(t_b) \quad \text{and} \quad B_n^K(a,b) = \frac{1}{n}\sum_{i=1}^{n} W_i(t_a)W_i(t_b),$$

for $a, b \in \{1, \dots, K\}$, and then set $R_n^K = L_n^K + B_n^K$.

We construct the smooth curves Y_i by setting $Y_i(t_j) = \sum_{a=1}^{r} c_{ia}\lambda_a^{1/2}\eta_a(t_j)$, where $\lambda_1, \dots, \lambda_r$ are positive scalars and $c_{ia} \sim N(0,1)$. We consider three different cases for the functions $\eta_1, \dots, \eta_r$ (which are, by construction, the eigenfunctions of $\mathscr{L}$). In the first case, we take the $\{\eta_j\}_{j=1}^r$ as the first r Fourier basis elements (denoted by FB in the sequel), and for the particular case $r = 1$, instead of using the constant function $\eta_1(t) = 1$, we take $\eta_1(t) = \sin(2\pi t)$; in the second case, the $\{\eta_j\}_{j=1}^r$ are constructed as the Gram–Schmidt orthogonalisation of the first r analytic functions (denoted by AC in the sequel) from the following list:

$$\begin{aligned} &\eta_1(t) = 5t\sin(2\pi t), \qquad \eta_2(t) = t\cos(2\pi t) - 3, \\ &\eta_3(t) = 5t + \sin(2\pi t) - 2, \\ &\eta_4(t) = \cos(4\pi t) + (t/2)^2, \qquad \eta_5(t) = \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} t(1-t). \end{aligned}$$

Finally, in the third case, we take the $\{\eta_j\}_{j=1}^r$ as the first r shifted Legendre polynomials $\tilde{P}_i(x)$ (denoted by LP in the sequel) defined as

$$\begin{aligned} &\eta_1(t) = 6t^2 - 6t + 1, \qquad \eta_2(t) = 2t - 1, \qquad \eta_3(t) = 1, \\ &\eta_4(t) = 20t^3 - 30t^2 + 12t - 1, \qquad \eta_5(t) = 70t^4 - 140t^3 + 90t^2 - 20t + 1. \end{aligned}$$

The rough curves W_i are produced in one of the following three ways:

1. We set $W_i(t_j) = \sum_{a=0}^{q} \theta_a \varepsilon_{i,j-a}$, where $q = \lceil K\delta/2 \rceil$, $\theta_0 = 1$, $\theta_1, \dots, \theta_q \in (-1, 1)$ are scalars and $\varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} N(0,1)$ (denoted by MA in the sequel).

TABLE 1
Scenarios for the simulation study

Scenarios	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Y_i	FB	AC	LP	FB	AC	LP	FB	AC	LP
W_i	MA	MA	MA	TRI	TRI	TRI	RBB	RBB	RBB

2. We set $W_i(t_j) = \sum_{a=1}^{d} b_{ia}\beta_a^{1/2}\psi_a(t_j)$, where $\beta_1, \dots, \beta_d$ are positive scalars and $b_{ia} \sim N(0,1)$. The functions ψ_a are triangular functions of norm 1 with support $[(a-1)\delta, a\delta]$ (denoted by TRI in the sequel).

3. We set $W_i(t_j) = \sum_{a=1}^{d} b_{ia}\beta_a^{1/2}\psi_a(t_j)$, where $\beta_1, \dots, \beta_d$ are positive scalars and $b_{ia} \sim N(0,1)$. The functions ψ_a are realisations of reflected Brownian bridges defined on $[(a-1)\delta, a\delta]$ (denoted by RBB in the sequel).

The nine different scenarios resulting from the three possible choices for the eigenfunctions η and the three possible choices for the rough component W are summarised in Table 1.

For each scenario, we consider 6 different combinations of the rank and bandwidth parameters r and δ, as given in the Table 2.

Finally, we also consider two different regimes for the choice of the eigenvalues $\lambda_1 < \cdots < \lambda_r$ of $\mathscr{L}$ and $\beta_1 < \cdots < \beta_d$ of $\mathscr{B}$; the first one can be seen as the easy case where there is a clear ordering distinction between the two sets, that is, $\lambda_r \gg \beta_1$ (regime 1); the second one is the interlaced case, when $\lambda_r < \beta_1 < \lambda_{r-1}$ (regime 2). In regime 1, the r eigenvalues λ are equally spaced between $\lambda_1 = 1.45$ and $\lambda_r = 0.25$, and we use $\lambda_1 = 0.25$ for $r = 1$. In regime 2, the eigenvalues $\{\lambda_1, \dots, \lambda_r\}$ are equally spaced between $\lambda_1 = 1$ and $\lambda_r = 0.04$. In either regime, the rough processes are simulated with $\beta_1 = 0.09$. The remaining eigenvalues for the scenarios (TRI) or (RBB) are smaller than 0.04 and decreasing toward zero, while those for the scenario (MA) are slowly decreasing toward zero, yielding a challenging situation in regime 2, since in this case there is more than one eigenvalue of the rough process that exceeds the smallest eigenvalue of the smooth process. For each combination (r, δ) with $r > 1$ of Table 2, we consider each of the two regimes and for the particular case $r = 1$, we consider only regime 1. In total, we consider 10 different cases in each one of the nine simulation scenarios.

TABLE 2
Different values of the rank and bandwidth parameter

Combination	**1**	**2**	**3**	**4**	**5**	**6**
r	1	1	3	3	5	5
δ	0.05	0.1	0.05	0.1	0.05	0.1

Our simulation study is divided into two parts. We first illustrate how the scree plots used to select the rank r of the operator $\mathscr{L}$ behave for the different scenarios. These show that using the scree plot as a basis for selection can be a very reasonable approach. We then compare our estimator $\hat{L}_n^K$ of L_n^K to the one obtained by three other methods: a direct use of a truncated Karhunen–Loève expansion; the spline smoothing approach popularised by Ramsay and Silverman [18]; and the PACE method of Yao et al. [21]. We also construct the estimated predictors $\hat{Y}_n^K$ of Y^K for a subset of the scenarios in order to probe their predictive accuracy. In doing this, we use the true rank of $\mathscr{L}$, as the simulations are computationally very intensive, and it would be infeasible to use an automatic selection method (and of course, it would be impossible to make a choice based on inspection of scree plots for all replications). Note that for the rest of this section we consider the maximal bandwidth of B^K to be 10 instead of $K/4 = 25$ (without emphasising it by a new notation), since one would rarely expect a rough process to have such a long memory, and since using a smaller maximal bandwidth value gives more stable and accurate numerical results. We have also carried out a simulation study to probe the performance of the estimators $\hat{L}_n^K$, $\hat{B}_n^K$ and $\hat{Y}_n^K$ when the data are corrupted by measurement errors and/or high frequency noise. The results can be found in the Supplementary Material ([8], Section 7.3), and are qualitatively very similar to those presented in the main text.

8.1. *Rank selection.* In order to probe the appropriateness of using a scree-type plot in order to estimate the rank r of the operator $\mathscr{L}$, we ran simulations on one sample of each scenario, each combination of the parameters r and δ and both regimes (for a total of $9 \times 6 + 9 \times 4 = 90$ simulations). As explained in Section 7, we plot the function $f(i) = \|P^K \circ (R_n^K - \hat{C}_i\hat{C}_i^\top)\|_F^2$, where $\hat{C}_i \in \mathbb{R}^{K\times i}$ is the minimiser of the optimisation problem (7.2), and then we select the rank j beyond which $f(j)$ levels out, that is, beyond which no meaningful reduction to the objective function is achieved. In practice we evaluate the function f over $i = 1, \dots, 10$ and not over $1, \dots, K/4 - 1 = 24$ as mentioned in the theory since the procedure is quite computationally intensive; it is clear from the resulting plots that this is not restrictive. The results are presented by scenario and by regime in Figure 2. Since the functions f are not on the same scale for every regime and every combination, we plotted a normalised version of f given by $f(i)/\|P^K \circ R_n^K\|_F^2$. For each scenario, the function f for the samples generated with $r = 5$ are in black, the ones generated with $r = 3$ are in red and the ones generated with $r = 1$ in blue. The dotted vertical lines indicate the location of the true rank, that is, 5 (in black), 3 (in red) and 1 (in blue). The figure reveals that for most of the scenarios, we would select the rank quite accurately in regime 1 and we would underestimate it a little bit in regime 2. In further simulations (reported in the Supplementary Material [8], Section 7.1) we study the effect of rank misspecification. It seems that underestimation is quite impactful in Regime 1 (noninterlaced eigenvalues) and that overestimation does not have a severe impact in both regimes, which suggests that one should not hesitate to over-estimate the rank relative to what the scree-plot indicates.

8.2. *Comparisons.* We investigate the performance of our estimator of L_n^K, alongside the three following methods:

1. The spline smoothing approach, popularised by Ramsay and Silverman [18]: compute $\widetilde{X}_i$, the smooth version of the observed curves X_i, by using B-spline smoothing; then define the estimator of L_n^K as $\hat{L}_{\mathrm{RS}}^K(a,b) = \frac{1}{n}\sum_{i=1}^n \widetilde{X}_i(t_a)\widetilde{X}_i(t_b)$;
2. The PACE method (Yao et al. [21]) described in Section 1: the estimator of L_n^K is given by $\hat{L}_{\mathrm{PACE}}^K(a,b) = \widetilde{\rho}(t_a,t_b)$. Of course it must be noted that PACE was primarily introduced for the sparse sampling case, but it can still be used in a dense setting.
3. Truncation of the empirical Karhunen–Loève (KL) expansion: we derive the spectral decomposition of R_n^K, and the estimator of L_n^K is simply equal to a spectrally truncated version thereof, at a level rk, where rk is chosen such that the variance explained is at least 95%.

For every choice of scenario (A)–(I), rank/bandwidth combination (1)–(6), and eigenvalue regime (regime 1 or regime 2), we simulate 100 replications for a sample size of $n = 300$ on a grid of $K = 100$ points. Results for different values of n and K can be found in the Supplementary Material [8], Section 7.2. For each replicate, we determine the estimators given by the four different methods, and calculate their normalised error, by evaluating the function $\mathrm{Err}(u) = (\|u - L_n^K\|_F)/\|L_n^K\|_F$ at every one of these estimators. We then form the ratio between our method's relative error (in the denominator) and the relative error of each of the three other methods (in the numerator). Consequently, we calculate 3×100 ratios per simulation regime. Their corresponding first quartiles, medians and third quartiles are presented in Table 3 (regime 1) and in Table 4 (regime 2), where those medians exceeding 1 have been highlighted in bold. These indicate settings where our approach typically performs comparably or at least as well any as the approach it is being compared to. Corresponding boxplots are provided in the Supplementary Material ([8], Section 7.4), allowing for a finer appreciation of the distribution of relative errors.

Of course, one cannot expect there to be a uniformly best method (for instance, the KL expansion is expected to perform best when all the eigenfunctions are approximately mutually orthogonal and the eigenvalues are not interlaced). That being said, Tables 3 and 4 reveal that our method has a performance that is typically better than or comparable to that of the best competitor in all but one scenarios/combinations. The exceptional case corresponds to a situation where the smooth curves were generated with the first 5 Legendre polynomials. In this particular setup, our optimisation problem was quite unstable due to the particular shape of the matrix L_n^K—it had very high values on the band relative to values

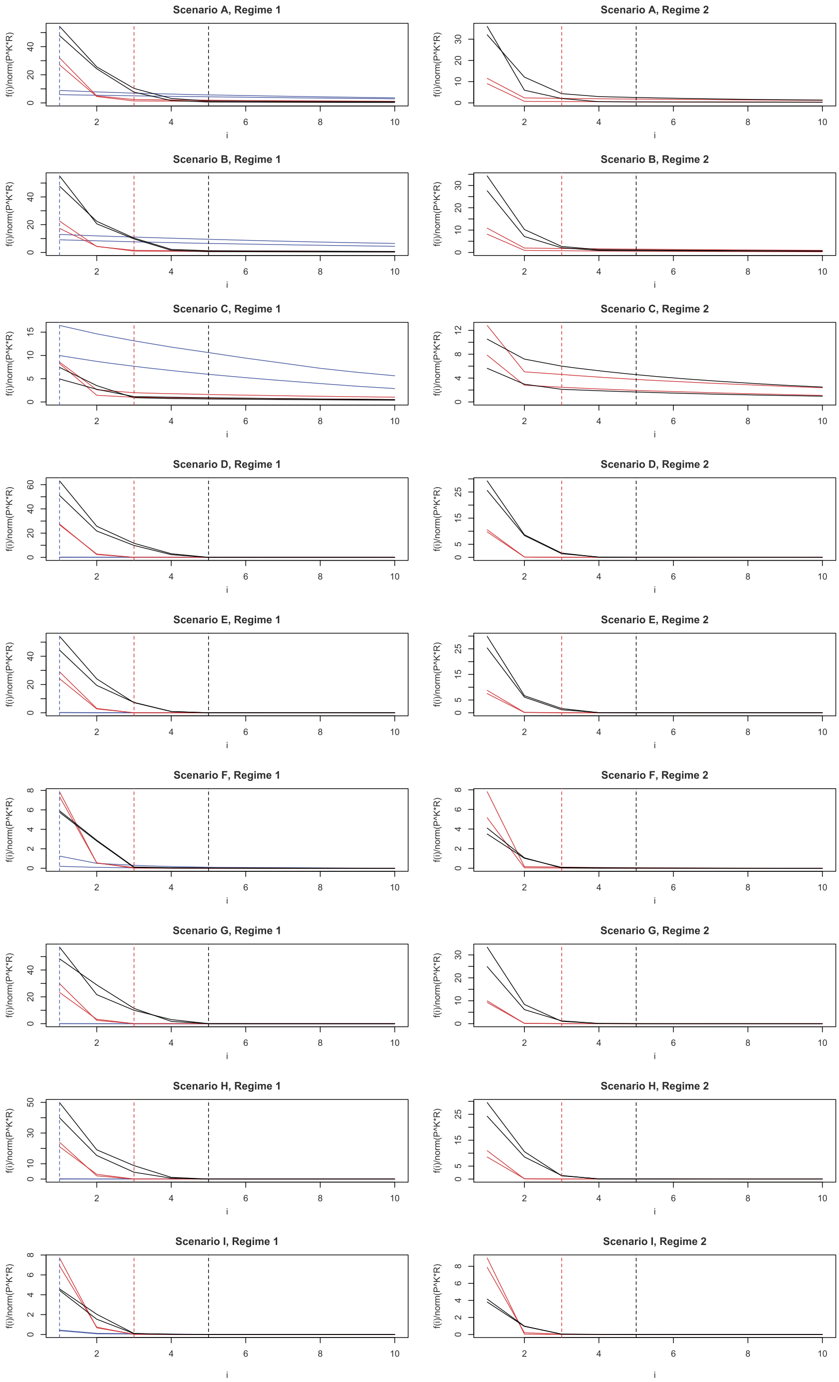

FIG. 2. *Plots of the function $f(\cdot)$ (defined in Section 7) normalised by $\|P^K \circ R_n^K\|_F^2$ for a given scenario, a given combination of parameters and a given regime. The curves in black correspond to a setting with $r = 5$, those in red to a setting with $r = 3$ and those in blue to a setting with $r = 1$.*

TABLE 3

Table containing the median (the first and third quartiles are in parentheses) of the ratios for the three methods we compared our method with and for the 9 *scenarios we considered with the regime* 1. *We highlight in bold the medians that exceed* 1

Scenario	**(rk, δ)**	**PACE**	**KL**	**RS**
		Regime 1		
A	(1, 0.05)	**4.01** (2.51, 6.46)	**2.87** (1.93, 4.18)	**4.15** (3.59, 5.16)
	(1, 0.10)	**4.44** (2.21, 7.83)	**3.40** (2.01, 5.48)	**4.92** (3.79, 6.03)
	(3, 0.05)	**3.19** (2.31, 4.60)	**2.89** (2.19, 3.73)	**3.02** (2.59, 3.46)
	(3, 0.10)	**3.10** (2.13, 4.50)	**2.75** (1.89, 3.97)	**2.89** (2.40, 3.32)
	(5, 0.05)	**2.58** (2.07, 3.26)	**2.41** (2.04, 2.92)	**2.04** (1.81, 2.33)
	(5, 0.10)	**2.20** (1.79, 2.86)	**2.10** (1.71, 2.60)	**1.87** (1.60, 2.08)
B	(1, 0.05)	**3.95** (2.05, 5.80)	**3.09** (1.79, 4.46)	**4.30** (3.51, 5.02)
	(1, 0.10)	**3.54** (1.83, 6.12)	**2.55** (1.55, 4.83)	**4.18** (3.44, 5.08)
	(3, 0.05)	**2.93** (2.55, 4.11)	**2.85** (2.36, 3.55)	**2.72** (2.37, 3.03)
	(3, 0.10)	**3.16** (2.49, 4.14)	**2.74** (2.22, 3.46)	**2.71** (2.43, 3.11)
	(5, 0.05)	**1.91** (1.49, 2.83)	**1.84** (1.48, 2.38)	**1.49** (1.23, 1.72)
	(5, 0.10)	**1.62** (1.28, 2.20)	**1.57** (1.25, 2.03)	**1.35** (1.07, 1.61)
C	(1, 0.05)	**2.22** (0.87, 4.20)	**1.05** (0.49, 2.27)	**2.82** (2.17, 3.71)
	(1, 0.10)	**1.34** (0.71, 3.02)	0.63 (3.38, 1.95)	**2.23** (1.01, 3.78)
	(3, 0.05)	**2.08** (1.58, 2.90)	**1.73** (1.28, 2.27)	**2.19** (1.78, 2.59)
	(3, 0.10)	**1.52** (1.08, 2.36)	**1.33** (0.79, 2.01)	**1.95** (1.33, 2.45)
	(5, 0.05)	0.43 (0.37, 0.55)	0.5 (0.48, 0.75)	0.42 (0.28, 0.74)
	(5, 0.10)	0.49 (0.40, 0.70)	0.51 (0.48, 0.69)	0.44 (0.28, 0.74)
D	(1, 0.05)	**11.7** (9.89, 12.8)	**11.7** (9.89, 12.8)	**10.5** (8.77, 11.6)
	(1, 0.10)	**21.0** (18.3, 26.5)	**21.9** (18.2, 26.4)	**16.1** (13.4, 19.3)
	(3, 0.05)	**6.83** (5.98, 7.41)	**6.66** (5.85, 7.33)	**5.00** (5.21, 6.46)
	(3, 0.10)	**11.2** (9.62, 12.9)	**10.8** (9.10, 12.4)	**8.80** (7.34, 10.0)
	(5, 0.05)	**4.51** (3.91, 5.18)	**4.27** (3.68, 4.95)	**3.92** (3.38, 4.52)
	(5, 0.10)	**7.50** (6.20, 8.65)	**7.11** (5.65, 8.24)	**5.94** (4.88, 6.74)
E	(1, 0.05)	**7.77** (6.97, 9.13)	**7.76** (6.97, 9.12)	**7.03** (6.17, 8.01)
	(1, 0.10)	**15.1** (12.6, 18.0)	**15.0** (12.6, 18.0)	**11.0** (9.41, 13.4)
	(3, 0.05)	**5.55** (5.05, 6.31)	**5.73** (5.15, 6.61)	**4.88** (4.45, 5.60)
	(3, 0.10)	**9.15** (7.81, 10.7)	**9.36** (8.00, 11.0)	**7.08** (5.98, 8.25)
	(5, 0.05)	**2.83** (2.26, 3.62)	**3.03** (2.39, 3.95)	**2.54** (1.95, 3.12)
	(5, 0.10)	**5.40** (4.31, 6.71)	**5.55** (4.56, 7.09)	**4.30** (3.34, 5.30)
F	(1, 0.05)	**8.91** (7.56, 10.2)	**9.05** (7.69, 10.3)	**7.78** (6.77, 9.08)
	(1, 0.10)	**18.2** (14.6, 24.5)	**18.3** (14.7, 24.6)	**13.3** (10.9, 17.9)
	(3, 0.05)	**5.43** (4.58, 6.31)	**5.67** (4.82, 6.67)	**4.69** (3.89, 5.51)
	(3, 0.10)	**9.84** (8.83, 11.2)	**10.2** (9.12, 11.5)	**7.47** (6.51, 8.43)
	(5, 0.05)	0.51 (0.18, 0.86)	0.52 (0.19, 0.91)	0.44 (0.15, 0.72)
	(5, 0.10)	**1.03** (0.47, 2.11)	**1.07** (0.49, 2.20)	0.73 (0.36, 1.57)

TABLE 3
(*Continued*)

Scenario	(rk, δ)	**PACE**	**KL**	**RS**
G	(1, 0.05)	**13.5** (10.2, 17.0)	**13.4** (10.2, 16.8)	**12.1** (9.43, 15.0)
	(1, 0.10)	**17.2** (13.0, 24.6)	**17.2** (13.0, 25.0)	**15.6** (11.6, 20.9)
	(3, 0.05)	**9.78** (8.17, 11.8)	**9.21** (7.38, 11.2)	**7.93** (6.97, 9.71)
	(3, 0.10)	**9.76** (7.94, 12.2)	**9.34** (7.58, 12.2)	**8.64** (7.00, 10.7)
	(5, 0.05)	**7.05** (6.07, 8.36)	**7.15** (5.93, 8.67)	**5.64** (4.85, 7.23)
	(5, 0.10)	**6.93** (5.68, 8.23)	**6.44** (5.37, 8.03)	**6.00** (5.04, 7.46)
H	(1, 0.05)	**11.0** (8.29, 13.8)	**10.9** (8.49, 13.8)	**9.29** (7.66, 11.8)
	(1, 0.10)	**14.2** (10.4, 18.2)	**14.2** (10.5, 18.2)	**11.7** (8.96, 16.0)
	(3, 0.05)	**7.76** (6.74, 9.89)	**8.72** (7.00, 10.2)	**6.89** (5.65, 7.85)
	(3, 0.10)	**8.67** (6.83, 11.2)	**8.63** (6.88, 11.3)	**7.95** (6.19, 10.2)
	(5, 0.05)	**4.80** (3.41, 6.20)	**6.01** (4.49, 8.14)	**4.03** (2.94, 5.44)
	(5, 0.10)	**5.36** (3.82, 6.89)	**5.60** (3.89, 7.17)	**4.67** (3.38, 5.95)
I	(1, 0.05)	**11.1** (9.31, 13.7)	**11.7** (9.68, 14.2)	**9.87** (8.21, 12.4)
	(1, 0.10)	**16.0** (11.4, 20.6)	**16.2** (11.5, 20.7)	**13.8** (9.87, 17.4)
	(3, 0.05)	**7.13** (6.00, 9.25)	**7.61** (6.49, 10.0)	**6.03** (5.21, 7.29)
	(3, 0.10)	**7.72** (6.29, 9.58)	**8.17** (6.49, 9.99)	**6.76** (5.46, 8.43)
	(5, 0.05)	**1.06** (0.65, 1.53)	**1.33** (0.72, 1.92)	0.88 (0.53, 1.27)
	(5, 0.10)	0.94 (0.18, 1.77)	0.99 (0.19, 1.82)	0.78 (0.15, 1.54)

outside the band, rendering matrix completion difficult. Consequently, some of the replications returned estimators that where completely off, as is indicated in the table by the small values of the first quartile for the scenarios C, F and I with $r = 5$. Of course, all the results need to be taken with a grain of salt, as we make use of the true rank when constructing our estimator, which in practice is unknown and must be selected (and of course, the methods to which we compare also involve the choice of tuning parameters, depending on which their performance may vary). These comparisons should thus be viewed as a benchmark, rather than a claim to superiority, as we compare to methods not specifically tailored for the problem at hand.

In practice, it may of course be that the rough component is indeed pure noise. In order to check whether our method performs comparably well with the other methods in this more classical setup, we additionally consider a scenario where the smooth curves are generated using a Fourier basis and the rough curves are discrete white noise. In this situation, the matrix B^K representing the discretised kernel b is precisely diagonal instead of just banded. The results are presented in the Table 5. Surprisingly, it appears that our method performs equally well or better than all other methods in all scenarios considered. A likely explanation is that, even when the process W has a diagonal kernel, its finite sample empirical kernel will not be exactly diagonal, but banded (since some empirical correlations will exist).

TABLE 4

Table containing the median (the first and third quartiles are in parentheses) of the ratios for the three methods we compared our method with and for the 9 *scenarios we considered with the regime* 2. *We highlight in bold the medians that exceed* 1

Scenario	**Combination**	**PACE**	**KL**	**RS**
		Regime 2		
A	(3, 0.05)	**1.84** (1.16, 2.54)	**1.87** (1.09, 2.79)	**2.26** (1.28, 2.86)
	(3, 0.10)	**1.20** (0.95, 1.87)	0.98 (0.83, 1.89)	**1.14** (0.78, 2.17)
	(5, 0.05)	**1.06** (0.87, 1.61)	0.96 (0.86, 1.72)	**1.08** (0.62, 1.76)
	(5, 0.10)	**1.01** (0.84, 1.24)	0.93 (0.82, 1.25)	0.91 (0.63, 1.22)
B	(3, 0.05)	**2.11** (1.29, 2.90)	**2.22** (1.22, 2.95)	**2.06** (1.30, 2.65)
	(3, 0.10)	**1.32** (1.05, 1.78)	**1.10** (0.91, 1.73)	**1.26** (0.73, 2.24)
	(5, 0.05)	0.94 (0.82, 1.10)	0.89 (0.80, 1.04)	0.75 (0.44, 1.07)
	(5, 0.10)	**1.04** (0.87, 1.24)	0.94 (0.80, 1.17)	0.90 (0.60, 1.33)
C	(3, 0.05)	**1.18** (0.88, 1.61)	0.80 (0.64, 1.44)	**1.25** (0.93, 2.02)
	(3, 0.10)	**1.15** (0.85, 1.62)	0.72 (0.58, 1.53)	**1.35** (0.83, 1.91)
	(5, 0.05)	0.68 (0.54, 0.89)	0.53 (0.48, 0.71)	0.79 (0.52, 1.32)
	(5, 0.10)	0.74 (0.54, 1.03)	0.56 (0.47, 1.04)	0.77 (0.58, 1.26)
D	(3, 0.05)	**5.70** (5.06, 6.62)	**5.59** (5.03, 6.65)	**4.93** (4.42, 5.73)
	(3, 0.10)	**10.7** (8.66, 12.2)	**10.5** (8.48, 12.2)	**8.03** (6.39, 9.37)
	(5, 0.05)	**3.58** (3.10, 4.18)	**3.48** (3.05, 4.03)	**3.08** (2.73, 3.59)
	(5, 0.10)	**6.81** (5.64, 8.09)	**6.63** (5.54, 7.72)	**5.27** (4.23, 6.17)
E	(3, 0.05)	**4.60** (3.89, 5.43)	**4.66** (3.96, 5.45)	**4.16** (3.60, 4.81)
	(3, 0.10)	**8.59** (6.96, 10.2)	**8.65** (7.00, 10.2)	**6.51** (5.22, 7.80)
	(5, 0.05)	**2.09** (1.11, 2.76)	**2.14** (1.13, 2.82)	**1.84** (0.94, 2.45)
	(5, 0.10)	**3.96** (3.15, 5.46)	**4.24** (3.33, 5.72)	**3.12** (2.42, 4.27)
F	(3, 0.05)	**1.13** (0.06, 2.74)	**1.17** (0.07, 2.83)	0.99 (0.06, 2.47)
	(3, 0.10)	**3.45** (0.16, 7.03)	**3.55** (0.16, 7.20)	**2.61** (0.11, 5.21)
	(5, 0.05)	0.78 (0.07, 1.43)	0.81 (0.07, 1.50)	0.66 (0.06, 1.27)
	(5, 0.10)	0.70 (0.09, 2.85)	0.71 (0.09, 2.95)	0.52 (0.07, 2.13)
G	(3, 0.05)	**7.87** (6.60, 9.69)	**7.31** (6.22, 9.55)	**6.56** (5.56, 8.07)
	(3, 0.10)	**8.05** (6.46, 9.91)	**8.02** (6.41, 9.92)	**7.03** (5.58, 9.10)
	(5, 0.05)	**5.73** (4.73, 6.52)	**7.03** (5.95, 8.53)	**4.94** (3.92, 5.68)
	(5, 0.10)	**5.87** (4.77, 7.88)	**5.75** (4.69, 7.92)	**5.30** (4.35, 7.00)
H	(3, 0.05)	**7.10** (6.07, 8.22)	**6.99** (5.73, 8.16)	**6.06** (5.13, 7.17)
	(3, 0.10)	**7.51** (6.03, 9.43)	**7.61** (6.09, 9.53)	**6.74** (5.63, 8.19)
	(5, 0.05)	**3.84** (3.16, 4.91)	**5.26** (4.11, 6.90)	**3.40** (2.64, 4.14)
	(5, 0.10)	**3.89** (1.76, 5.46)	**4.30** (1.82, 5.84)	**3.53** (1.47, 5.02)
I	(3, 0.05)	**4.94** (3.27, 6.13)	**5.32** (3.48, 6.54)	**4.41** (3.12, 5.30)
	(3, 0.10)	**3.11** (0.20, 6.11)	**3.16** (0.20, 6.24)	**2.87** (0.17, 5.12)
	(5, 0.05)	0.59 (0.06, 1.47)	0.67 (0.07, 1.58)	0.49 (0.05, 1.24)
	(5, 0.10)	**1.16** (0.14, 2.54)	**1.20** (0.15, 2.60)	**1.02** (0.11, 2.38)

TABLE 5

Table containing the median (*the first and third quartiles are in parentheses*) *of the ratios for the three methods we compared our method with for the classical scenario where the rough component is a white noise. We highlight in bold the results that exceed* 1

r	**PACE**	**KL**	**RS**
		Regime 1	
1	**1.92** (1.69, 2.16)	**1.76** (1.53, 2.05)	**4.08** (3.77, 4.33)
3	**2.90** (2.58, 3.16)	**3.02** (2.66, 3.28)	**3.36** (3.05, 3.53)
5	**2.80** (2.61, 3.01)	**2.78** (2.56, 3.02)	**2.40** (2.24, 2.65)
		Regime 2	
3	**1.63** (1.45, 1.76)	**2.01** (1.85, 2.19)	**2.36** (2.22, 2.57)
5	**1.28** (1.16, 1.37)	**1.48** (1.36, 1.61)	**1.64** (1.45, 1.75)

8.3. *Prediction of the smooth curves.* We selected 6 different cases in order to probe the performance of our estimated predictor $\hat{Y}_n^K$ as a proxy for the true predictor $\Pi(X^K)$. We considered, for both regimes, combination 5 of scenario A, combination 4 of scenarios F and combination 6 of scenario H. For every sample, we calculated the average of the approximation of the normalised mean integrated squared error of $\hat{Y}_n^K$:

$$\mathrm{relMISE} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j=1}^{K}[\hat{Y}_{n,i}^K(t_j) - \Pi(X_i^K)(t_j)]^2}{\sum_{j=1}^{K}[\Pi(X_i^K)(t_j)]^2}.$$

Figure 3 contains boxplots of their distributions. These illustrate that, as expected, our predictions perform better when the eigenvalues of $\mathscr{L}$ and $\mathscr{B}$ are not interlaced.

9. Proofs of formal statements.

9.1. *Proofs of theorems in Section 3.*

PROOF OF THEOREM 1. Since the eigenfunctions of $\mathscr{L}_1$ and $\mathscr{L}_2$ are analytic and $\max\{r_1, r_2\} < \infty$, it follows that the corresponding covariance kernels are bivariate analytic functions on $[0, 1]^2$ ([13], Theorem 4.3.3).

This being the case, the zero set of either kernel is at most 1-dimensional, unless the kernels are uniformly zero ([13], Theorem 6.33). Since our theorem follows trivially if $\mathscr{L}_1$ and $\mathscr{L}_2$ are the zero operator, we can assume that their kernels are not uniformly zero. Thus, if we can show that the two kernels coincide on an open subset U of $[0, 1]^2$, then they will necessarily coincide everywhere on $(0, 1)^2$, and thus on $[0, 1]^2$ by continuity. This, in particular, will in turn imply that $\mathscr{B}_1$ and $\mathscr{B}_2$ also coincide.

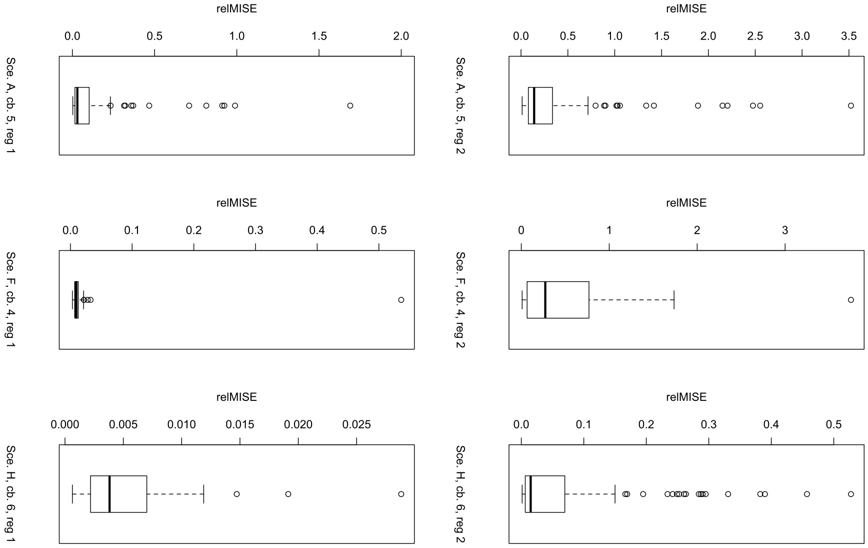

FIG. 3. *Distributions of* relMISE. *First row*: *scenario A with* $r = 5, \delta = 0.05$, *regime* 1 *on the left and regime* 2 *on the right. Middle row*: *scenario F with* $r = 3, \delta = 0.1$, *regime* 1 *on the left and regime* 2 *on the right. Last row*: *scenario H with* $r = 5, \delta = 0.1$, *regime* 1 *on the left and regime* 2 *on the right.*

Without lost of generality, assume that $\delta_1 \geq \delta_2$. Define

$$U = (\delta_1, 1) \times (0, 1 - \delta_1).$$

Since $\mathscr{L}_1 + \mathscr{B}_1 = \mathscr{L}_2 + \mathscr{B}_2$, but $\mathscr{B}_1 = \mathscr{B}_2 = 0$ on U, it must be that the kernels of $\mathscr{L}_1$ and $\mathscr{L}_2$ coincide on the open set U, and the proof is complete. □

The proof of Proposition 1 can be found in the Supplementary Material [8], Section 5. Moving on, the proof of Theorem 2 rests upon the observation that it is essentially a statement regarding matrix completion. Our strategy of proof will thus be to translate our functional conditions on $\mathscr{B}$ and $\mathscr{L}$ into matrix properties of L^K and B^K that suffice for unique matrix completion. We first develop the said matrix properties in the form of Lemma 1 and Theorem 4.

LEMMA 1. *Let $b(s,t)$ be a continuous kernel on $[0,1]^2$ such that $b(s,t) = 0$ whenever $|s - t| > \delta$, and let $(t_1, \dots, t_K) \in \mathcal{T}_K$ be a grid of K points. Then the matrix $B^K = \{b(t_i, t_j)\}_{i,j=1}^K$ is banded with bandwidth $2\lceil \delta \cdot K \rceil + 1$.*

THEOREM 4. *Let $\mathscr{L}$ have kernel $\ell(s,t) = \sum_{i=1}^r \lambda_i \eta_i(s) \eta_i(t)$ with $r < \infty$ and real analytic orthonormal eigenfunctions $\{\eta_1, \dots, \eta_r\}$. If $K > r$, then the minors of order r of the matrix $L^K = \{\ell(t_i, t_j)\}_{i,j=1}^K$ are all nonzero, almost everywhere on $\mathcal{T}_K$.*

PROOF. First, notice that from $\ell(s,t)=\sum_{i=1}^{r}\lambda_i\eta_i(s)\eta_i(t)$, we have

$$L_{jl}^{K}=\sum_{i=1}^{r}\lambda_i\eta_i(t_j)\eta_i(t_l).$$

Thus, L^K can be written as $U^K\Sigma(U^K)^\top$, where

$$U^K=\begin{pmatrix}\eta_1(t_1) & \eta_2(t_1) & \cdots & \eta_r(t_1)\\ \eta_1(t_2) & \eta_2(t_2) & \cdots & \eta_r(t_2)\\ \vdots & \vdots & & \vdots\\ \eta_1(t_K) & \eta_2(t_K) & \cdots & \eta_r(t_K)\end{pmatrix} \quad\text{and}$$

$$\Sigma=\begin{pmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \lambda_r\end{pmatrix}. \tag{9.1}$$

Any $r\times r$ submatrix of L^K obtained by deleting rows and columns, can then be written as

$$U_F^K\Sigma(U_{F'}^K)^\top,$$

where U_F^K (resp., $U_{F'}^K$) is an $r\times r$ matrix obtained by deleting rows of U^K whose indices are not included in $F\subseteq\{1,\dots,K\}$ (resp., F'). The condition that any minor of order r of L^K be nonzero is then equivalent to the condition that

$$\det[U_F^K\Sigma(U_{F'}^K)^\top]=\det[U_F^K]\det[\Sigma]\det[U_{F'}^K]\neq 0,$$

for any subset F, $F'\subseteq\{1,\dots,K\}$ of cardinality r. By construction $\det(\Sigma)\neq 0$, so the minor condition is then equivalent to requiring that $\det(U_F^K)\neq 0$ for any subset $F\subseteq\{1,\dots,K\}$ of cardinality r.

We will show that this is indeed the case almost everywhere on $\mathcal{T}_K$. Let μ denote Lebesgue measure on $\mathcal{T}_K$ and let $F=\{1,\dots,r\}$, without loss of generality (so that U_F^K is formed by keeping the first r rows of U^K). Using the Leibniz formula, we have that $\det(U_F^K)$ can be written as the function

$$D(t_1,\dots,t_r)=\sum_{\sigma\in S_r}\varepsilon(\sigma)\prod_{i=1}^{r}\eta_i(t_{\sigma(i)}),$$

where S_r is the symmetric group on r elements and $\varepsilon(\sigma)$ is the signature of the permutation σ. Note that the function D is real analytic on $(0,1)^r$, by virtue of each η_i being real analytic on $(0,1)$.

We will now proceed by contradiction. Assume that

$$\mu\{(x_1,\dots,x_K)\in\mathcal{T}_K : D(x_1,\dots,x_r)=0\}>0.$$

Since μ is Lebesgue measure, it follows that the Hausdorff dimension of the set $A = \{(x_1, \dots, x_r) : D(x_1, \dots, x_r) = 0\}$ is equal to r. However, since D is analytic, Krantz and Parks [13], Theorem 6.33, implies the dichotomy: either D is constant everywhere on $(0, 1)^r$, or the set A is at most of dimension $r - 1$. Thus, it must be that D is everywhere constant on $(0, 1)^r$, the constant being of course zero:

$$D(x_1, \dots, x_r) = \sum_{\sigma \in S_r} \varepsilon(\sigma) \prod_{i=1}^{r} \eta_i(x_{\sigma(i)}) = 0 \qquad \forall (x_1, \dots, x_r) \in (0, 1)^r.$$

Now fix $(x_1, \dots, x_{r-1})$ and apply to D (viewed as a function of x_r only) the continuous linear functional $T_{\eta_r}(f) = \langle f, \eta_r \rangle$. We obtain that for all $(x_1, \dots, x_{r-1}) \in (0, 1)^r$:

$$\begin{aligned} 0 = \langle D, \eta_r \rangle &= \sum_{\sigma \in S_r} \varepsilon(\sigma) \biggl[\prod_{i : \sigma(i) \neq r} \eta_i(x_{\sigma(i)}) \biggr] \langle \eta_{\sigma^{-1}(r)}, \eta_r \rangle \\ &= \sum_{\sigma \in S_{r-1}} \varepsilon(\sigma) \prod_{i=1}^{r-1} \eta_i(x_{\sigma(i)}). \end{aligned}$$

Applying iteratively the continuous linear functionals $T_{\eta_j}(f) = \langle f, \eta_j \rangle$ to D while keeping $(x_1, \dots, x_{j-1})$ fixed then leads to

$$\eta_1(y) = 0 \qquad \forall y \in (0, 1).$$

This last equality contradicts the fact that η_1 is of norm one, and allows us to conclude that $\mu\{(x_1, \dots, x_K) \in \mathcal{T}_K : D(x_1, \dots, x_r) = 0\} = 0$. □

We now prove Theorem 2 by demonstrating that the matrix properties of (L^K, B^K) that derive from its assumptions are sufficient for unique matrix completion. The proof is inspired by Proposition 2.12 of [12].

PROOF OF THEOREM 2. Given our conditions, Lemma 1 implies that $B_1, B_2 \in \mathbb{R}^{K \times K}$ are banded matrices with bandwidth $2\lceil \delta_i \cdot K \rceil + 1$, for $i \in \{1, 2\}$.

Let $\delta = \max\{\delta_1, \delta_2\}$ and assume without loss of generality that $r_1 \geq r_2$. Let Ω be the set of indices on which both B_1 and B_2 vanish, which by Lemma 1 is $\Omega = \{(i, j) \in \{1, \dots, K\}^2 : |i - j| > \lceil \delta \cdot K \rceil\}$. From $L_1 + B_1 = L_2 + B_2$, we obtain that $\{L_1\}_{ij} = \{L_2\}_{ij}, \forall (i, j) \in \Omega$. Let Ω_A be the set of indices of a submatrix formed by the first r_1 rows and the last r_1 columns of a $K \times K$ matrix, the condition $K \geq K^* = \frac{2r_1+2}{1-2\delta}$ implies that $\Omega_A \subset \Omega$, which in turn implies that the matrices L_1 and L_2 contain a common submatrix A of dimension $r_1 \times r_1$.

Assume that all minors of order r_1 of L_1 are nonzero. Then the determinant of A is nonzero, which implies that the rank of L_2 is also r_1. We thus establish that L_1 and L_2 are two rank r_1 matrices equal on Ω. Let L^* be a matrix equal to L_1 on Ω, but unknown at those indices that do not belong to Ω. We will now show that there exists a unique rank r_1 completion of L^*. Due to the band pattern of the unobserved entries of L^* and the inequality $K \geq K^* = \frac{2r_1+2}{1-2\delta}$, it is possible to find a submatrix of L^* of dimension $(r_1 + 1) \times (r_1 + 1)$ with only one unobserved entry, denoted x^*. Using the fact that the determinant of any square submatrix of dimension larger than $r_1 + 1$ is zero, we obtain a linear equation of the form $ax^* + b = 0$, where a is equal to the determinant of a submatrix of dimension $r_1 \times r_1$. Since we assume that any minor of order r_1 is nonzero, we have that $a \neq 0$ and the previous equation has a unique solution. It is then possible to impute the value of x^*. Applying this procedure iteratively until all missing entries are determined allows us to uniquely complete the matrix L^* into a rank r_1 matrix. In summary, we have demonstrated that when all minors of order r_1 of L_1 are nonzero, it holds that $L^* = L_1 = L_2$ and hence $B_1 = B_2$. Theorem 4 assures us that L_1 indeed has nonvanishing minors of order r_1 almost everywhere on $\mathcal{T}_K$, and so we conclude that it must be that $L_1 = L_2$ and $B_1 = B_2$ almost everywhere on $\mathcal{T}_K$. □

9.2. *Proofs of theorems in Section 4.*

PROOF OF PROPOSITION 2. Since $\delta < 1/4$ and $K \geq 4r+1$ implies $K \geq \frac{2r+2}{1-2\delta}$, Theorem 2 implies that the objective function (4.1) achieves its minimal value of r at L^K. To elaborate, note that any minimiser of (4.1) must equal L^K on the set $\Omega = \{(i, j) \in \{1, \dots, K\}^2 : |i - j| > \lceil \delta \cdot K \rceil\}$, as it has to satisfy the constraint $\|P^K(R^K - \theta)\|_F^2 = 0$. Consequently, any minimiser has a nonzero minor of order r in Ω, implying that its rank is bounded below by r. Thus its rank must be exactly r, since L^K satisfies the constraint and has rank r. We conclude that any minimiser of (4.1) must be equal to L^K everywhere, following the same iterative completion process as in the second part of the proof of Theorem 2 (see immediately above).

We now turn to prove that $L^K = \arg\min_{\theta \in \mathbb{R}^{K \times K}}\{\|P^K \circ (R^K - \theta)\|_F^2 + \tau \operatorname{rank}(\theta)\}$, for all $\tau > 0$ sufficiently small. Since we have established that L^K uniquely solves

$$\min_{\theta \in \mathbb{R}^{K \times K}} \operatorname{rank}\{\theta\} \quad \text{subject to} \quad \|P^K \circ (R^K - \theta)\|_F^2 = 0,$$

it follows that for all $\tau > 0$ and any $\theta \in \mathbb{R}^{K \times K}$ of rank greater or equal to r, we have that

$$\|P^K \circ (R^K - L^K)\|_F^2 + \tau \operatorname{rank}(L^K) < \|P^K \circ (R^K - \theta)\|_F^2 + \tau \operatorname{rank}(\theta).$$

We thus concentrate on matrices $\theta \in \mathbb{R}^{K \times K}$ of rank at most $r - 1$, for $r > 1$. Let

$$\mu = \min_{\theta \in \mathbb{R}^{K \times K}, \operatorname{rank}(\theta) \leq r-1} \{\|P^K \circ (R^K - \theta)\|_F^2\} > 0.$$

Now let $\tau_* = \frac{\mu}{r-1}$. Then, for any $\tau < \tau_*$, and any θ of rank less than r,

$$\begin{aligned} \|P^K \circ (R^K - L^K)\|_F^2 + \tau \operatorname{rank}(L^K) &= \tau r < \mu + \tau \\ &\leq \|P^K \circ (R^K - \theta)\|_F^2 + \tau \operatorname{rank}(\theta). \end{aligned}$$

In summary, putting our results together, we have shown that for all $\tau \in (0, \tau_*)$,

$$L^K = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{K \times K}} \{ \| P^K \circ (R^K - \theta) \|_F^2 + \tau \operatorname{rank}(\theta) \}.$$

Finally, it is worth pointing out that although τ_* depends on r, this does not mean that the objective function depends on unknowns: r can be shown (using Theorem 4) to be equal to the rank of the submatrix formed by the first $\lceil K/4 \rceil$ rows and the last $\lceil K/4 \rceil$ columns of R^K, and thus we can determine τ_* directly from the matrix R^K. This completes the proof. □

9.3. *Proofs of theorems in Section 6.*

PROOF OF THEOREM 3. We begin by the usual bias/variance decomposition

$$\begin{aligned} \|\hat{\mathscr{L}}_n^K - \mathscr{L}\|_{\text{HS}}^2 &\le 2\|\hat{\mathscr{L}}_n^K - \mathscr{L}^K\|_{\text{HS}}^2 + 2\|\mathscr{L}^K - \mathscr{L}\|_{\text{HS}}^2 \\ &= 2K^{-2}\|\hat{L}_n^K - L^K\|_{\text{F}}^2 + 2\|\mathscr{L}^K - \mathscr{L}\|_{\text{HS}}^2. \end{aligned}$$

For the second term (bias), we note that by a Taylor expansion

$$\begin{aligned} &\int_0^1 \int_0^1 (\ell(x, y) - \ell_K(x, y))^2 \, dx \, dy \\ &\quad = \sum_{i,j=1}^K \int_{I_{i,K}} \int_{I_{j,K}} (\ell(x, y) - \ell(t_i, t_j))^2 \, dx \, dy \\ &\quad \le \sum_{i,j=1}^K \int_{I_{i,K}} \int_{I_{j,K}} 2K^{-2} \sup_{(x,y) \in I_{i,K} \times I_{j,K}} \|\nabla \ell(x, y)\|_2^2 \\ &\quad \le 2K^{-2} \sup_{(x,y) \in [0,1]^2} \|\nabla \ell(x, y)\|_2^2. \end{aligned}$$

Without loss of generality, we assume that the data are rescaled so that $K^{-1}\text{trace}(R_n^K) = 1$. To show that $K^{-2}\|\hat{L}_n^K - L^K\|_{\text{F}}^2 = O_{\mathbb{P}}(n^{-1})$ almost everywhere on $\mathcal{T}_K$, define Θ_K to be the space of $K \times K$ nonnegative matrices of trace at most K. Consider the functionals

$$\mathbb{S}_{n,K} : \Theta_K \to [0, \infty), \qquad \mathbb{S}_{n,K}(\theta) = \underbrace{K^{-2}\|P^K \circ (\theta - R_n^K)\|_{\text{F}}^2}_{\mathbb{M}_{n,K}(\theta)} + \tau \operatorname{rank}(\theta),$$

$$S_K : \Theta_K \to [0, \infty), \qquad S_K(\theta) = \underbrace{K^{-2}\|P^K \circ (\theta - R^K)\|_{\text{F}}^2}_{M_K(\theta)} + \tau \operatorname{rank}(\theta),$$

where $P_K(i, j) = \mathbf{1}\{|i - j| > \lceil K/4 \rceil\}$. Note that, since $K \ge 4r + 4$, Theorem 2 implies that for almost all grids, L^K is the unique minimiser of S_K, for all $\tau > 0$ sufficiently small. From now on, fix such a grid, and let $\tau > 0$ be sufficiently small.

First, we will show that $\hat{L}_n^K$ is consistent for L^K. To this aim, note that

$$\begin{aligned}\left|\mathbb{S}_{n,K}(\theta) - S_K(\theta)\right| &= \left|\mathbb{M}_{n,K}(\theta) - M_K(\theta)\right| \\ &= K^{-2}\left|\|P^K \circ (\theta - R_n^K)\|_F^2 - \|P^K \circ (\theta - R^K)\|_F^2\right| \\ &\le K^{-2}\left|\|P^K \circ (\theta - R_n^K)\|_F - \|P^K \circ (\theta - R^K)\|_F\right| \\ &\quad \times \left(\|P^K \circ (\theta - R_n^K)\|_F + \|P^K \circ (\theta - R^K)\|_F\right) \\ &\le K^{-2}\|P^K \circ (R_n^K - R^K)\|_F \left(2\|\theta\|_F + \|R_n^K\|_F + \|R^K\|_F\right).\end{aligned}$$

It follows that $\sup_{\theta\in\Theta_K}\left|\mathbb{S}_{n,K}(\theta) - S_K(\theta)\right| \overset{n\to\infty}{\to} 0$ almost surely, and given that $S_K(\theta)$ is lower semicontinuous with a unique minimum at L^K, and $\hat{L}_n^K \in \Theta_K$, consistency of $\hat{L}_n^K$ for L^K follows [19, Corollary 3.2.3].

Next we show that $\mathrm{rank}(\hat{L}_n^K)$ is consistent for the true rank. Suppose that this is not true. Then there exist $\epsilon > 0$, $\delta > 0$ and a subsequence $\{n_j\}$ such that $\mathbb{P}\{|\mathrm{rank}(\hat{L}_{n_j}^K) - r| > \epsilon\} > \delta$ for all $j \ge 1$. So, $\mathbb{P}\{\mathrm{rank}(\hat{L}_{n_j}^K) \neq r\} > \delta$ for all $j \ge 1$. Thus, there exist possibly two subsequences $\{j_l\}$ and $\{k_l\}$ such that $\mathbb{P}\{\mathrm{rank}(\hat{L}_{j_l}^K) > r\} > \delta/2$ and $\mathbb{P}\{\mathrm{rank}(\hat{L}_{k_l}^K) < r\} > \delta/2$ for all $l \ge 1$. The latter possibility is impossible since $\hat{L}_n^K$ is consistent, and matrices of rank at most $r-1$ form a closed set. For the first possibility, since $\hat{L}_{j_l}^K$ converges to L^K in probability, there exists a further subsequence $\{j_{l_m}\}$ such that $\mathrm{rank}(\hat{L}_{j_{l_m}}^K) > r$ for all $m \ge 1$ and $\hat{L}_{j_{l_m}}^K$ converges to L^K as $m \to \infty$. Without any loss of generality, we can assume that $P(\mathrm{rank}(\hat{L}_{j_{l_m}}^K) > r) > \delta/2$ for all $m \ge 1$, and $\hat{L}_{j_{l_m}}^K$ converges to L^K as $m \to \infty$ almost surely (or take further subsequences). So, the set where both of these events hold has probability at least $\delta/2$. Working on this set, and by $\hat{L}_{j_{l_m}}^K$ being a minimiser,

$$\begin{aligned}&\mathbb{M}_{n,K}(\hat{L}_{j_{l_m}}^K) + \tau(r+1) \\ &\quad = K^{-2}\|P^K \circ (\hat{L}_{j_{l_m}}^K - R_n^K)\|_F^2 + \tau(r+1) \\ &\quad \le K^{-2}\|P^K \circ (\hat{L}_{j_{l_m}}^K - R_n^K)\|_F^2 + \tau\,\mathrm{rank}(\hat{L}_{j_{l_m}}^K) \\ &\quad \le \inf_{\theta\in\Theta_K:\mathrm{rank}(\theta)=r}\{K^{-2}\|P^K \circ (\theta - R_n^K)\|_F^2 + \tau\mathrm{rank}(\theta)\} \\ &\quad = \inf_{\theta\in\Theta_K:\mathrm{rank}(\theta)=r} K^{-2}\|P^K \circ (\theta - R_n^K)\|_F^2 + \tau r \\ &\quad = \inf_{\theta\in\Theta_K:\mathrm{rank}(\theta)=r} \mathbb{M}_{n,K}(\theta) + \tau r,\end{aligned} \tag{9.2}$$

for all $m \ge 1$. But $\sup_{\theta\in\Theta_K}|\mathbb{M}_{n,K}(\theta) - M_K(\theta)| \to 0$ almost surely, so $\mathbb{M}_{n,K}(\hat{L}_{j_{l_m}}^K) - M_K(\hat{L}_{j_{l_m}}^K) \to 0$. Also, by continuity, $M_K(\hat{L}_{j_{l_m}}^K) \to M_K(L^K) = 0$.

Consequently, $\mathbb{M}_{n,K}(\hat{L}^K_{j_{l_m}}) \to 0$. Now note that, on the set $\{\theta \in \Theta_K : \operatorname{rank}(\theta) = r\}$, the sequence of functions $\mathbb{M}_{n,K}(\theta)$ are equi-Lipschitz continuous almost surely. So, from the uniform convergence, we will also have

$$\inf_{\theta \in \Theta_K:\operatorname{rank}(\theta)=r} \mathbb{M}_{n,K}(\theta) \to \inf_{\theta \in \Theta_K:\operatorname{rank}(\theta)=r} M_K(\theta) = 0.$$

Combining the above facts and using (9.2), we arrive at the contradiction that $\tau \leq 0$. Summarising, if we define

$$d^2(\theta, L^K) = K^{-2}\|\theta - L^K\|_F^2 + \tau\big|\operatorname{rank}(\theta) - \operatorname{rank}(L^K)\big|,$$

then we have $d(\hat{L}^K_n, L^K) \to 0$ in probability as $n \to \infty$. We will now use consistency in conjunction with [19, Theorem 3.4.1], to obtain the rate. Write

$$\Delta(\theta) = S_K(\theta) - S_K(L^K) = K^{-2}\|P^K \circ (\theta - L^K)\|_{\mathrm{F}}^2 + \tau(\operatorname{rank}(\theta) - r).$$

Choose $\eta^2 < \tau$ and observe that, for any θ with $\operatorname{rank}(\theta) \neq r$, we must have $d^2(\theta, L^K) \geq \tau|\operatorname{rank}(\theta) - r| \geq \tau > \eta^2$, which implies that $d(\theta, L^K) > \eta$. Thus, no matrix θ with $\operatorname{rank}(\theta) \neq r$ satisfies $\gamma/2 < d(\theta, L^K) < \gamma$ for $\gamma < \eta$. Hence,

$$\inf_{\theta \in \Theta_K:\gamma/2<d(\theta,L^K)<\gamma} \Delta(\theta) = \inf_{\theta \in \Theta_K:\gamma/2<d(\theta,L^K)<\gamma,\operatorname{rank}(\theta)=r} \Delta(\theta).$$

We will show that the latter quantity is bounded below by $\alpha_0\gamma^2$, where $\alpha_0 > 0$ and $\gamma < \eta$, for $\eta > 0$ sufficiently small. This is equivalent to showing that

$$\inf_{\theta \in \Theta_K:\gamma^2/4<\|\theta-L^K\|_F^2<\gamma^2,\operatorname{rank}(\theta)=r} \|P^K \circ (\theta - L^K)\|_F^2 > \alpha_1\gamma^2, \tag{9.3}$$

for some $\alpha_1 > 0$. We argue by contradiction. Fix any θ with $\operatorname{rank}(\theta) = r$ and $\|\theta - L^K\|_F^2 > d$, where we write $d = \gamma^2/4$ for tidiness. Suppose that $\|P^K \circ (\theta - L^K)\|_F^2 < \beta d$, for some $\beta \in (0, 1/2)$. Now, we can always write $\theta = L^K + A + B$, where $A = P^K \circ A$ and $P^K \circ B = 0$ [simply define $A = P^K \circ (\theta - L^K)$ and $B = \theta - L^K - A$]. If $\|P^K \circ (\theta - L^K)\|_F^2 < \beta d$ for some $\beta \in (0, 1/2)$, we have $\|A\|_F^2 < \beta d$ and $\|A + B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 > d$. So, $\|B\|_F^2 > (1 - \beta)d > d/2$ and there exists an element (j, k) (in the band defined by P^K) such that $|B_{j,k}| > \sqrt{d/(2c_K)}$, where c_K is the total number of elements in the band. Observe that $\theta_{j,k} = L^K_{j,k} + B_{j,k}$.

Now, we know that all possible minors of L^K of order r are nonzero, and for sufficiently small η, the same is true in an η-neighbourhood of L^K, which includes θ. Let the indices of the rows and columns of such an $r \times r$ sub-matrix of L^K, say C_K, be denoted by $\{p_1, p_2, \dots, p_r\}$ and $\{q_1, q_2, \dots, q_r\}$, respectively. Exploiting the structure of the band, choose this sub-matrix in such a way that the sub-matrix elements and the entries $\{(j, q_l) : 1 \leq l \leq r\}$ and $\{(p_l, k) : 1 \leq l \leq r\}$ lie outside the band defined by P^K. Consider the sub-matrix of order r of θ, say E, by taking the same rows and columns as in C_K. Define the sub-matrix F (resp. D) of order $(r+$

1) obtained by adjoining to E (resp. to C_K), the elements $\mathbf{q}_1 = (\theta_{j,q_1}, \dots, \theta_{j,q_r})'$, $\mathbf{q}_2 = (\theta_{p_1,k}, \dots, \theta_{p_r,k})'$ and $\theta_{j,k}$ [resp. the elements $\mathbf{c}_1 = (L^K_{j,q_1}, \dots, L^K_{j,q_r})'$, $\mathbf{c}_2 = (L^K_{p_1,k}, \dots, L^K_{p_r,k})'$ and $L^K_{j,k}$]. So,

$$F = \begin{bmatrix} \theta_{j,k} & \mathbf{q}_1' \\ \mathbf{q}_2 & E \end{bmatrix} \quad \text{and } D = \begin{bmatrix} L^K_{j,k} & \mathbf{c}_1' \\ \mathbf{c}_2 & C_K \end{bmatrix}.$$

Then, for η sufficiently small, we have that

$$|B_{j,k}| = |\mathbf{q}_1' E^{-1} \mathbf{q}_2 - \mathbf{c}_1' C_K^{-1} \mathbf{c}_2| < \kappa \| P^K \circ (\theta - L^K)\|_F < \kappa\sqrt{\beta d},$$

by the fact that the map $(\mathbf{q}_1, \mathbf{q}_2, E) \mapsto \mathbf{q}_1' E^{-1} \mathbf{q}_2$ is locally Lipschitz at any $(\mathbf{c}_1, \mathbf{c}_2, C_K)$ as constructed above. So for β chosen to be sufficiently small, we have contradicted the fact that $|B_{j,k}| > \sqrt{d/(2c_K)}$. In summary, for some $\beta \in (0, 1/2)$ sufficiently small, we must have $\|P^K \circ (\theta - L^K)\|_F^2 > \beta d$ if θ is a rank r matrix with $\|\theta - L^K\|_F^2 > d$, as sought.

Next, define

$$\begin{aligned} D(\theta) &= \mathbb{S}_{n,K}(\theta) - S_K(\theta) - \mathbb{S}_{n,K}(L^K) + S_K(L^K) \\ &= \mathbb{M}_{n,K}(\theta) - M_K(\theta) - \mathbb{M}_{n,K}(L^K) + M_K(L^K). \end{aligned}$$

We expand $(\mathbb{M}_{n,K} - M_K)$ in a first-order Taylor expansion with Lagrange remainder, around L^K, which gives for a certain $\tilde{p} \in [0, 1]$ and $\tilde{\theta} = \tilde{p} L^K + (1 - \tilde{p})\theta$:

$$\begin{aligned} D(\theta) &= \langle \mathbb{M}_{n,K}'(\tilde{\theta}), \theta - L^K \rangle_{\mathrm{F}} - \langle M_K'(\tilde{\theta}), \theta - L^K \rangle_{\mathrm{F}} \\ &= K^{-2} \langle 2P^K \circ (\tilde{\theta} - R_n^K), \theta - L^K \rangle_{\mathrm{F}} - K^{-2} \langle 2P^K \circ (\tilde{\theta} - R^K), (\theta - L^K) \rangle_{\mathrm{F}} \\ &= K^{-2} \langle 2P^K \circ \tilde{\theta} - 2P^K \circ \tilde{\theta} - 2P^K \circ R_n^K + 2P^K \circ R^K, \theta - L^K \rangle_{\mathrm{F}} \\ &\le K^{-2} \| 2P^K \circ (R_n^K - R^K)\|_{\mathrm{F}} \|\theta - L^K\|_{\mathrm{F}} \\ &\le 2K^{-1} \|R_n^K - R^K\|_{\mathrm{F}} K^{-1} \|\theta - L^K\|_{\mathrm{F}}. \end{aligned}$$

Since $\mathbb{E}\|X\|_{L^2}^4 < \infty$, the process $X(s)X(t)$ is trace class on $[0, 1]^2$, and thus has a continuous covariance kernel on $[0, 1]^4$ (and consequently a continuous variance function on $[0, 1]^2$). Assume without loss of generality that $\mathbb{E}X = 0$. Since the observations $X_i(t_j)$ are independent for distinct i, and since $X_m(t_j)X_m(t_j)$ is an unbiased estimator of $\mathbb{E}[X(t_j)X(t_j)]$, we have

$$\begin{aligned} & K^{-2}\mathbb{E}\|R_n^K - R^K\|_F^2 \\ &\quad = \sum_{i=1}^{K} \sum_{j=1}^{K} K^{-2} \mathbb{E}\left[\frac{1}{n} \sum_{m=1}^{n} X_m(t_{i,K}) X_m(t_{j,K}) - \mathbb{E}[X(t_{i,K}) X(t_{j,K})]\right]^2 \\ &\quad = \frac{K^{-2}}{n} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathrm{Var}[X(t_{i,K}) X(t_{j,K})] \le \frac{1}{n} \sup_{(s,t) \in [0,1]^2} \mathrm{Var}[X(s)X(t)] = \frac{C}{n}, \end{aligned}$$

and $C = \sup_{[0,1]^2} \mathrm{Var}[X(s)X(t)] < \infty$. Once again, by the choice of η in relation to τ, it follows that

$$\begin{aligned}\mathbb{E}\Bigl\{\sup_{\theta\in\Theta_K : d_n(\theta,L^K)<\gamma} \bigl|D(\theta)\bigr|\Bigr\} &= \mathbb{E}\Bigl\{\sup_{\theta\in\Theta_K : d_n(\theta,L^K)<\gamma,\mathrm{rank}(\theta)=r} \bigl|D(\theta)\bigr|\Bigr\}\\ &= \mathbb{E}\Bigl\{\sup_{\theta\in\Theta_K : K^{-1}\|\theta-L^K\|_{\mathrm{F}}<\gamma} \bigl|D(\theta)\bigr|\Bigr\}\\ &\le 2\gamma K^{-1}\mathbb{E}\|R_n^K - R^K\|_{\mathrm{F}}\\ &\le 2\gamma\sqrt{\frac{C}{n}}.\end{aligned}$$

It now follows [19, Theorem 3.4.1] that if $\hat{L}_n^K$ is an approximate minimiser of $\mathbb{S}_{n,K}$, in the sense given by the assumptions, then it holds that

$$nd_n^2(\hat{L}_n^K, L^K) = nK^{-2}\|\hat{L}_n^K - L^K\|_F^2 + n\tau\bigl|\mathrm{rank}(\hat{L}_n^K) - r\bigr| = O_{\mathbb{P}}(1),$$

from which we conclude that

$$K^{-2}\|\hat{L}_n^K - L^K\|_F^2 = O_{\mathbb{P}}(1) \quad \text{and} \quad \bigl|\mathrm{rank}(\hat{\mathscr{L}}_n^K) - r\bigr| = O_{\mathbb{P}}(n^{-1}).$$

Finally, we now turn our attention to the estimated eigenfunctions. Since these are finitely many, we will omit the index indicating the order of an eigenfunction for tidiness, and consider an eigenfunction η. Let η^K be the K-resolution step function approximation of η, $\eta^K(x) = \sum_{j=1}^K \eta(t_{j,K})\mathbf{1}\{x \in I_{j,K}\}$. Then, by Taylor expanding,

$$\begin{aligned}\int_0^1 (\eta(x) - \eta^K(x))^2\,dx &= \sum_{j=1}^K \int_{I_{j,K}} (\eta(x) - \eta(t_{j,K}))^2\,dx\\ &\le \sum_{j=1}^K \int_{I_{j,K}} K^{-2}\|\eta'\|_\infty^2 = \frac{\|\eta'\|_\infty^2}{K^2}.\end{aligned}$$

It follows that

$$\begin{aligned}\|\hat{\eta} - \eta\|_{L^2}^2 &\le 2\|\hat{\eta} - \eta^K\|_{L^2}^2 + 2\|\eta^K - \eta\|_{L^2}^2\\ &\le c\|\hat{\mathscr{L}}_n^K - \mathscr{L}^K\|_{\mathrm{HS}}^2 + \frac{2\|\eta'\|_\infty^2}{K^2} = O_{\mathbb{P}}(n^{-1}) + \frac{2\|\eta'\|_\infty^2}{K^2}.\end{aligned}$$

The constant c can be chosen uniformly over the order of eigenfunction, since there are only $r < \infty$ eigenfunctions to consider. The convergence rate for $\sup_j |\hat{\lambda}_j - \lambda_j|$ follows from the inequality $\sup_j |\hat{\lambda}_j - \lambda_j| \le \|\hat{\mathscr{L}}_n^K - \mathscr{L}\|_{\mathrm{HS}}$ (Bosq [2], equation (4.43)).

□

The proofs of Corollaries 1 and 3 can be found in the Supplementary Material [8], Section 5. Corollary 2 follows directly from Theorem 3 and Corollary 1.

10. Concluding remarks. We conclude the paper with a short discussion and some perspectives regarding the role of smoothing, and the impact of high frequency noise and/or pure measurement error.

To smooth or not to smooth. As discussed in detail in Section 2 of the Supplementary Material, smoothing should be avoided prior to separating the smooth and rough components of the process, as it can confound the two types of variation and distort further analysis when $\mathscr{B}$ is not purely diagonal. At the same time, even if $\mathscr{B}$ is purely diagonal, our simulation results in Table 5 show that our method can still perform at least as well as classical smoothing-based methods, leading to no apparent loss in efficiency. Therefore, it seems that smoothing prior to separation is either not advisable, or not necessary. Smoothing *can* be applied, however, as a post-processing step, to each of the smooth and rough covariances obtained *after* our methodology has been applied (see the discussion at the end of Section 4). Such a post-processing smoothing step can lead to visually more appealing estimators of the smooth covariance $\mathscr{L}$; and, in the case of the rough covariance $\mathscr{B}$, to potentially more efficient estimators, if more regularity can be assumed on $\mathscr{B}$. In summary, we do not advocate that smoothing should be altogether replaced by our method. Instead, we suggest that in the presence of nondiagonal error covariance, smoothing is preferable as a post-processing rather than a pre-processing step. The two steps (separation and smoothing) are best seen as complementary.

High frequency noise. Our model $X = Y + W$ implicitly assumes that any high frequency fluctuations in X should be attributed to local variations due to W (i.e., rough components of variation exhibit short-range dependence). This reflects a common principle that high frequency features usually are localised in nature, as one assumes in many wavelet-based methods. Nevertheless, one may ask what may happen if there exist high frequency fluctuations in X that are *global*, that is, have analytic eigenfunctions, and so must be attributed to Y—for example, cases where Y is not precisely of finite rank, but has most of its variation expressed in r eigenfunctions, and a small part of its variation expressed by higher order eigenfunctions. This residual variation can be considered as nuisance noise, but one may wonder if it would impact the performance of our method. Simulations carried out in Section 7 of the Supplementary Material [8] consider precisely this scenario, by adding higher frequency components to Y, such as high frequency trigonometric functions or diffusion processes with analytic eigenfunctions. It is observed that the presence of this high frequency noise *has a negligible effect* on the performance of our method, at least as far as estimation of $\mathscr{L}$ is concerned. Estimation of $\mathscr{B}$ is more appreciably affected, since the band is now contaminated, and more structural knowledge would be required to reliably separate the global from the local high frequency fluctuations. More detailed discussion of this point can be found in the Supplementary Material [8], Section 7.

Pure measurement error. It can happen that further to the rough – yet trace-class – component W, there is still some i.i.d. measurement error which enters the model at the level of discrete measurement. The presence of such measurement does not impact the method of estimation of $\mathscr{L}$, since this is based on removing a band of size $\lceil K/4 \rceil$ from the empirical covariance R_n^K, and carrying out matrix completion. Without additional assumptions, however, we would not be able to estimate the kernel b of $\mathscr{B}$ near the diagonal. Additional simulations in Section 7 of the Supplementary Material consider contamination by pure measurement error, and corroborate these theoretical predictions.

Acknowledgments. We gratefully acknowledge support from the Swiss National Science Foundation. Victor M. Panaretos wishes to thank Prof. Hans-Georg Müller for stimulating discussions and Prof. Tilmann Gneiting for his helpful comments. We are grateful to Prof. Anirvan Chakraborty for several useful comments and suggestions. We also warmly thank three anonymous reviewers and two associate editors for their detailed, thoughtful, and constructive comments.

SUPPLEMENTARY MATERIAL

Supplement to "Functional data analysis by matrix completion" (DOI: 10.1214/17-AOS1590SUPP; .pdf). This supplement contains: additional theoretical results, a data analysis, proofs omitted from the main article, additional simulation results and further discussion.

REFERENCES

[1] Bauschke, H. H. and Borwein, J. M. (1996). On projection algorithms for solving convex feasibility problems. *SIAM Rev.* **38** 367–426. MR1409591

[2] Bosq, D. (2000). *Linear Processes in Function Spaces*: *Theory and Applications*. *Lecture Notes in Statistics* **149**. Springer, New York. MR1783138

[3] Bosq, D. (2014). Computing the best linear predictor in a Hilbert space. Applications to general ARMAH processes. *J. Multivariate Anal.* **124** 436–450. MR3147336

[4] Chen, Y. and Wainwright, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. Available at arXiv:1509.03025.

[5] Coleman, T. F. and Li, Y. (1994). On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. *Math. Program.* **67** 189–224. MR1305566

[6] Coleman, T. F. and Li, Y. (1996). An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.* **6** 418–445. MR1387333

[7] Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivariate Anal.* **12** 136–154. MR0650934

[8] Descary, M.-H. and Panaretos, V. M. (2018). Supplement to "Functional data analysis by matrix completion." DOI:10.1214/17-AOS1590SUPP.

[9] Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365

[10] Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer, New York. MR2920735

[11] HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, Chichester. MR3379106

[12] KIRÁLY, F. and TOMIOKA, R. (2012). A combinatorial algebraic approach for the identifiability of low-rank matrix completion. In *Proceedings of the* 29*th International Conference on Machine Learning*.

[13] KRANTZ, S. G. and PARKS, H. R. (2002). *A Primer of Real Analytic Functions*, 2nd ed. Birkhäuser, Boston, MA. MR1916029

[14] LI, Y. and HSING, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38** 3321–3351. MR2766854

[15] MATLAB (2012). Version 8.0.0.783 (R2012b). The MathWorks Inc., Natick, MA.

[16] OPSOMER, J., WANG, Y. and YANG, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.* **16** 134–153. MR1861070

[17] PANARETOS, V. M. and TAVAKOLI, S. (2013). Cramér–Karhunen–Loève representation and harmonic principal component analysis of functional time series. *Stochastic Process. Appl.* **123** 2779–2807. MR3054545

[18] RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

[19] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. Springer, New York. MR1385671

[20] WANG, J. L., CHIOU, J. M. and MÜLLER, H.-G. (2015). Review of functional data analysis. Available at arXiv:1507.05135.

[21] YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561

[22] YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106

INSTITUT DE MATHÉMATIQUES
ECOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
1015 LAUSANNE
SWITZERLAND
E-MAIL: marie-helene.descary@epfl.ch
victor.panaretos@epfl.ch