The Annals of Statistics
2012, Vol. 40, No. 5, 2483–2510
DOI: 10.1214/12-AOS1033

SAMPLED FORMS OF FUNCTIONAL PCA IN REPRODUCING KERNEL HILBERT SPACES

BY ARASH A. AMINI AND MARTIN J. WAINWRIGHT[1]

University of California, Berkeley

We consider the sampling problem for functional PCA (fPCA), where the simplest example is the case of taking time samples of the underlying functional components. More generally, we model the sampling operation as a continuous linear map from $\mathcal{H}$ to $\mathbb{R}^m$, where the functional components to lie in some Hilbert subspace $\mathcal{H}$ of L^2, such as a reproducing kernel Hilbert space of smooth functions. This model includes time and frequency sampling as special cases. In contrast to classical approach in fPCA in which access to entire functions is assumed, having a limited number m of functional samples places limitations on the performance of statistical procedures. We study these effects by analyzing the rate of convergence of an M-estimator for the subspace spanned by the leading components in a multi-spiked covariance model. The estimator takes the form of regularized PCA, and hence is computationally attractive. We analyze the behavior of this estimator within a nonasymptotic framework, and provide bounds that hold with high probability as a function of the number of statistical samples n and the number of functional samples m. We also derive lower bounds showing that the rates obtained are minimax optimal.

1. Introduction. The statistical analysis of functional data, commonly referred to as functional data analysis (FDA), is an established area of statistics with a great number of practical applications; see the books [26, 27] and references therein for various examples. When the data is available as finely sampled curves, say in time, it is common to treat it as a collection of continuous-time curves or functions, each being observed in totality. These datasets are then termed "functional," and various statistical procedures applicable in finite dimensions can be extended to this functional setting. Among such procedures is principal component analysis (PCA), which is the focus of present work.

If one thinks of continuity as a mathematical abstraction of reality, then treating functional data as continuous curves is arguably a valid modeling device. However, in practice, one is faced with finite computational resources and is forced to implement a (finite-dimensional) approximation of true functional procedures by some sort of truncation procedure, for instance, in the frequency domain. It is then

Received September 2011; revised April 2012.
[1]Supported in part by NSF Grants CCF-0545862 and DMS-09-07632.
MSC2010 subject classifications. Primary 62G05; secondary 62H25, 62H12, 41A35, 41A25.
Key words and phrases. Functional principal component analysis, time sampling, Fourier truncation, linear sampling operator, reproducing kernel Hilbert space.

important to understand the effects of this truncation on the statistical performance of the procedure. In other situations, such as in longitudinal data analysis [13], a continuous curve model is justified as a hidden underlying generating process to which one has access only through sparsely sampled measurements in time, possibly corrupted by noise. Studying how the time-sampling affects the estimation of the underlying functions in the presence of noise shares various common elements with the frequency-domain problem described above.

The aim of this paper is to study effects of "sampling"—in a fairly general sense—on functional principal component analysis in smooth function spaces. In order to do so, we adopt a functional-theoretic approach by treating the sampling procedure as a (continuous) linear operator. This set-up provides us with a notion of sampling general enough to treat both the frequency-truncation and time-sampling within a unified framework. We take as our smooth function space a Hilbert subspace $\mathcal{H}$ of $L^2[0,1]$ and denote the sampling operator by $\Phi : \mathcal{H} \to \mathbb{R}^m$. We assume that there are functions $x_i(t)$, $t \in [0,1]$, in $\mathcal{H}$ for $i = 1, \ldots, n$, generated i.i.d. from a probabilistic model (to be discussed). We then observe the collection $\{\Phi x_i\}_{i=1}^n \subset \mathbb{R}^m$ in noise. We refer to the index n as the number of *statistical samples*, and to the index m as the number of *functional samples*.

We analyze a natural M-estimator which takes the form of a regularized PCA in $\mathbb{R}^m$, and provide nonasymptotic bounds on the estimation error in terms of n and m. The eigen-decay of two operators govern the rates, the product of the sampling operator Φ and its adjoint, and the product of the map embedding $\mathcal{H}$ in L^2 and its adjoint. Our focus will be on the setting where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), in which case the two eigen-decays are intimately related through the kernel function $(s,t) \mapsto \mathbb{K}(s,t)$. In such cases, the two components of the rate interact and give rise to optimal values for the number of functional samples (m) in terms of the number of statistical samples (n) or vice versa. This has practical appeal in cases where obtaining either type of samples is costly.

Our model for the functions $\{x_i\}$ is an extension to function spaces of the *spiked covariance model* introduced by Johnstone and his collaborators [18, 19], and studied by various authors (e.g., [1, 19, 23]). We consider such models with r components, each lying within the Hilbert ball $\mathbb{B}_{\mathcal{H}}(\rho)$ of radius ρ, with the goal of recovering the r-dimensional subspace spanned by the spiked components in this functional model. We analyze our M-estimators within a high-dimensional framework that allows both the number of statistical samples n and the number of functional samples m to diverge together. Our main theoretical contributions are to derive nonasymptotic bounds on the estimation error as a function of the pair (m,n), which are shown to be sharp (minimax-optimal). Although our rates also explicitly track the number of components r and the smoothness parameter ρ, we do not make any effort to obtain optimal dependence on these parameters.

The general asymptotic properties of PCA in function spaces have been investigated by various authors (e.g., [8, 11, 15]). Accounting for smoothness of

functions by introducing various roughness/smoothness penalties is a standard approach, used in the papers [7, 24, 28, 29], among others. The problem of principal component analysis for sampled functions, with a similar functional-theoretic perspective, is discussed by Besse and Ramsey [5] for the noiseless case. A more recent line of work is devoted to the case of functional PCA with noisy sampled functions [10, 16, 32]. Cardot [10] considers estimation via spline-based approximation, and derives MISE rates in terms of various parameters of the model. Hall et al. [16] study estimation via local linear smoothing, and establish minimax-optimality in certain settings that involve a fixed number of functional samples. Both papers [10, 16] demonstrate trade-offs between the numbers of statistical and functional samples; we refer the reader to Hall et al. [16] for an illuminating discussion of connections between FDA and LDA approaches (i.e., having full versus sampled functions), which inspired much of the present work. We note that the regularization present in our M-estimator is closely related to classical roughness penalties [28, 29] in the special case of spline kernels, although the discussion there applies to fully-observed functions, as opposed to the sampled models considered here.

After initial posting of this work, we became aware of more recent work on sampled functional PCA. Working within the framework of Hall et al. [16], the analysis of Li and Hsing [21] allows for more flexible sample sizes per curve; they derive optimal uniform (i.e., L^∞) rates of convergence for local linear smoothing estimators of covariance function and the resulting eigenfunctions. Another line of work [17, 25] has analyzed sampled forms of Silverman's criterion [29], with some variations. Huang et al. [17] derive a criterion based on rank-one approximation coupled with scale invariance considerations, combined with an extra weighting of the covariance matrix. Xi and Zhao [25] also show the consistency of their estimator for both regular and irregular sampling. The regular (time) sampling setup in both papers have an overlap with our work; the eigenfunctions are assumed to lie in a second order Sobolev space, corresponding to a special case of a RKHS. However, even in this particular case, our estimator is different, and it is an interesting question whether a version of the results presented here can be used to show the minimax optimality of these Silverman-type criteria. There has also been recent work with emphasis on sampled functional covariance estimation, including the work of Cai and Yuan [9], who analyze an estimator which can be described as regularized least-squares with penalty being the norm of tensor product of RKHS with itself. They provide rates of convergence for the covariance function, from which certain rates (argued to be optimal within logarithmic factors) for eigenfunctions follow.

As mentioned above, our sampled model resembles very much that of spiked covariance model for high-dimensional principal component analysis. A line of work on this model has treated various types of sparsity conditions on the eigenfunctions [1, 19, 23]; in contrast, here the smoothness condition on functional components translates into an ellipsoid condition on the vector principal components. Perhaps

an even more significant difference is that in this paper, the effective scaling of noise in $\mathbb{R}^m$ is substantially smaller in some cases (e.g., the case of time sampling). This difference could explain why the difficulty of "high-dimensional" setting is not observed in such cases as one lets $m, n \to \infty$. On the other hand, a difficulty particular to our sampled model is the lack of orthonormality between components after sampling. It not only leads to identifiability issues, but also makes recovering individual components difficult.

In order to derive nonasymptotic bounds on our M-estimator, we exploit various techniques from empirical process theory (e.g., [30]), as well as the concentration of measure (e.g., [20]). We also exploit recent work [22] on the localized Rademacher complexities of unit balls in a reproducing kernel Hilbert space, as well as techniques from nonasymptotic random matrix theory, as discussed in Davidson and Szarek [12], in order to control various norms of random matrices. These techniques allow us to obtain finite-sample bounds that hold with high probability, and are specified explicitly in terms of the pair (m, n), and the underlying smoothness of the Hilbert space.

The remainder of this paper is organized as follows. Section 2 is devoted to background material on reproducing kernel Hilbert spaces, adjoints of operators, as well as the class of sampled functional models that we study in this paper. In Section 3, we describe M-estimators for sampled functional PCA, and discuss various implementation details. Section 4 is devoted to the statements of our main results, and discussion of their consequences for particular sampling models. In subsequent sections, we provide the proofs of our results, with some more technical aspects deferred to the supplementary material [3]. Section 5 is devoted to bounds on the subspace-based error. We conclude with a discussion in Section 6. In the supplementary material [3], Section 7 is devoted to proofs of bounds on error in the function space, whereas Section 8 provides proofs of matching lower bounds on the minimax error, showing that our analysis is sharp.

Notation. We will use $|||\cdot|||_{\mathrm{HS}}$ to denote the Hilbert–Schmidt norm of an operator or a matrix. The corresponding inner product is denoted as $\langle\langle \cdot, \cdot \rangle\rangle$. If T is an operator on a Hilbert space $\mathcal{H}$ with an orthonormal basis $\{e_j\}$, then $|||T|||_{\mathrm{HS}}^2 = \sum_j \|Te_j\|_{\mathcal{H}}^2$. For a matrix $A = (a_{ij})$, we have $|||A|||_{\mathrm{HS}}^2 = \sum_{i,j} |a_{ij}|^2$. For a linear operator Φ, the adjoint is denoted as Φ^*, the range as $\mathrm{Ra}(\Phi)$ and the kernel as $\mathrm{Ker}(\Phi)$.

2. Background and problem set-up. In this section, we begin by introducing background on reproducing kernel Hilbert spaces, as well as linear operators and their adjoints. We then introduce the functional and observation model that we study in this paper, and conclude with discussion of some approximation-theoretic issues that play an important role in parts of our analysis.

2.1. *Reproducing kernel Hilbert spaces*. We begin with a quick overview of some standard properties of reproducing kernel Hilbert spaces; we refer the reader

to the books [14, 31] and references therein for more details. A reproducing kernel Hilbert space (or RKHS for short) is a Hilbert space $\mathcal{H}$ of functions $f : T \to \mathbb{R}$ that is equipped with a symmetric positive semidefinite function $\mathbb{K} : T \times T \to \mathbb{R}$, known as the kernel function. We assume the kernel to be continuous, and the set $T \subset \mathbb{R}^d$ to be compact. For concreteness, we think of $T = [0, 1]$ throughout this paper, but any compact set of $\mathbb{R}^d$ suffices. For each $t \in T$, the function $R_t := \mathbb{K}(\cdot, t)$ belongs to the Hilbert space $\mathcal{H}$ and it acts as the *representer of evaluation*, meaning that $\langle f, R_t \rangle_{\mathcal{H}} = f(t)$ for all $f \in \mathcal{H}$.

The kernel $\mathbb{K}$ defines an integral operator $\mathcal{T}_{\mathbb{K}}$ on $L^2(T)$, mapping the function f to the function $g(s) = \int_T \mathbb{K}(s, t) f(t)\, dt$. By the spectral theorem in Hilbert spaces, this operator can be associated with a sequence of eigenfunctions $\psi_k, k = 1, 2, \ldots,$ in $\mathcal{H}$, orthogonal in $\mathcal{H}$ and orthonormal in $L^2(T)$, and a sequence of nonnegative eigenvalues $\mu_1 \geq \mu_2 \geq \cdots$. Most useful for this paper is the fact that any function $f \in \mathcal{H}$ has an expansion in terms of these eigenfunctions and eigenvalues, namely

$$f = \sum_{k=1}^{\infty} \sqrt{\mu_k} \alpha_k \psi_k \tag{1}$$

for some $(\alpha_k) \in \ell^2$. In terms of this expansion, we have the representations $\|f\|_{\mathcal{H}}^2 = \sum_{k=1}^{\infty} \alpha_k^2$ and $\|f\|_{L^2}^2 = \sum_{k=1}^{\infty} \mu_k \alpha_k^2$. Many of our results involve the decay rate of these eigenvalues: in particular, for some parameter $\alpha > 1/2$, we say that the kernel operator has eigenvalues with *polynomial-α decay* if there is a constant $c > 0$ such that

$$\mu_k \leq \frac{c}{k^{2\alpha}} \qquad \text{for all } k = 1, 2, \ldots. \tag{2}$$

Let us consider an example to illustrate.

EXAMPLE 1 (Sobolev class with smoothness $\alpha = 1$). In the case $T = [0, 1]$ and $\alpha = 1$, we can consider the kernel function $\mathbb{K}(s, t) = \min\{s, t\}$. As discussed in Appendix A of the supplementary material [3], this kernel generates the class of functions

$$\mathcal{H} := \big\{ f \in L^2\big([0, 1]\big) \mid f(0) = 0,\ f \text{ absolutely continuous and } f' \in L^2\big([0, 1]\big) \big\}.$$

The class $\mathcal{H}$ is an RKHS with inner product $\langle f, g \rangle_{\mathcal{H}} = \int_0^1 f'(t) g'(t)\, dt$, and the ball $\mathbb{B}_{\mathcal{H}}(\rho)$ corresponds to a Sobolev space with smoothness $\alpha = 1$. The eigendecomposition of the kernel integral operator is

$$\mu_k = \left[\frac{(2k - 1)\pi}{2} \right]^{-2}, \qquad \psi_k(t) = \sqrt{2} \sin\big(\mu_k^{-1/2} t\big), \qquad k = 1, 2, \ldots. \tag{3}$$

Consequently, this class has polynomial decay with parameter $\alpha = 1$.

We note that there are natural generalizations of this example to $\alpha = 2, 3, \ldots$, corresponding to the Sobolev classes of α-times differentiable functions; for example, see the books [4, 14, 31].

In this paper, the operation of generalized sampling is defined in terms of a bounded linear operator $\Phi : \mathcal{H} \to \mathbb{R}^m$ on the Hilbert space. Its adjoint is a mapping $\Phi^* : \mathbb{R}^m \to \mathcal{H}$, defined by the relation $\langle \Phi f, a \rangle_{\mathbb{R}^m} = \langle f, \Phi^* a \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $a \in \mathbb{R}^m$. In order to compute a representation of the adjoint, we note that by the Riesz representation theorem, the jth coordinate of this mapping—namely, $f \mapsto [\Phi f]_j$—can be represented as an inner product $\langle \phi_j, f \rangle_{\mathcal{H}}$, for some element $\phi_j \in \mathcal{H}$, and we can write

$$\Phi f = [\, \langle \phi_1, f \rangle_{\mathcal{H}} \quad \langle \phi_2, f \rangle_{\mathcal{H}} \quad \cdots \quad \langle \phi_m, f \rangle_{\mathcal{H}} \,]^T . \tag{4}$$

Consequently, we have $\langle \Phi f, a \rangle_{\mathbb{R}^m} = \sum_{j=1}^m a_j \langle \phi_j, f \rangle_{\mathcal{H}} = \langle \sum_{j=1}^m a_j \phi_j, f \rangle_{\mathcal{H}}$, so that for any $a \in \mathbb{R}^m$,

$$\Phi^* a = \sum_{j=1}^m a_j \phi_j. \tag{5}$$

This adjoint operator plays an important role in our analysis.

2.2. *Functional model and observations.* Let $s_1 \geq s_2 \geq s_3 \geq \cdots \geq s_r > 0$ be a fixed sequence of positive numbers, and let $\{f_j^*\}_{j=1}^r$ be a fixed sequence of functions orthonormal in $L^2[0, 1]$. Consider a collection of n i.i.d. random functions $\{x_1, \ldots, x_n\}$, generated according to the model

$$x_i(t) = \sum_{j=1}^r s_j \beta_{ij} f_j^*(t) \qquad \text{for } i = 1, \ldots, n, \tag{6}$$

where $\{\beta_{ij}\}$ are i.i.d. $N(0, 1)$ across all pairs (i, j). This model corresponds to a finite-rank instantiation of functional PCA, in which the goal is to estimate the span of the unknown eigenfunctions $\{f_j^*\}_{j=1}^r$. Typically, these eigenfunctions are assumed to satisfy certain smoothness conditions; in this paper, we model such conditions by assuming that the eigenfunctions belong to a reproducing kernel Hilbert space $\mathcal{H}$ embedded within $L^2[0, 1]$; more specifically, they lie in some ball in $\mathcal{H}$,

$$\| f_j^* \|_{\mathcal{H}} \leq \rho, \qquad j = 1, \ldots, r. \tag{7}$$

For statistical problems involving estimation of functions, the random functions might only be observed at certain times $(t_1, \ldots, t_m)$, such as in longitudinal data analysis, or we might collect only projections of each x_i in certain directions, such as in tomographic reconstruction. More concretely, in a *time-sampling model*, we observe m-dimensional vectors of the form

$$y_i = [\, x_i(t_1) \quad x_i(t_2) \quad \cdots \quad x_i(t_m) \,]^T + \sigma_0 w_i \qquad \text{for } i = 1, 2, \ldots, n, \tag{8}$$

where $\{t_1, t_2, \ldots, t_m\}$ is a fixed collection of design points, and $w_i \in \mathbb{R}^m$ is a noise vector. Another observation model is the *basis truncation model* in which we observe the projections of f onto the first m basis functions $\{\psi_j\}_{j=1}^m$ of the kernel operator—namely,

$$y_i = [\,\langle \psi_1, x_i\rangle_{L^2} \quad \langle \psi_2, x_i\rangle_{L^2} \quad \cdots \quad \langle \psi_m, x_i\rangle_{L^2}\,]^T + \sigma_0 w_i \qquad \text{for } i = 1, 2, \ldots, n, \tag{9}$$

where $\langle \cdot, \cdot \rangle_{L^2}$ represents the inner product in $L^2[0, 1]$.

In order to model these and other scenarios in a unified manner, we introduce a linear operator Φ_m that maps any function x in the Hilbert space to a vector $\Phi_m(x)$ of m samples, and then consider the linear observation model

$$y_i = \Phi_m(x_i) + \sigma_m w_i \qquad \text{for } i = 1, 2, \ldots, n. \tag{10}$$

This model (10) can be viewed as a functional analog of the spiked covariance models introduced by Johnstone [18, 19] as an analytically-convenient model for studying high-dimensional effects in classical PCA.

Both the time-sampling (8) and frequency truncation (9) models can be represented in this way, for appropriate choices of the operator Φ_m. Recall representation (4) of Φ_m in terms of the functions $\{\phi_j\}_{j=1}^m$.

- For the time sampling model (8), we set $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$, so that by the reproducing property of the kernel, we have $\langle \phi_j, f\rangle_{\mathcal{H}} = f(t_j)/\sqrt{m}$ for all $f \in \mathcal{H}$, and $j = 1, 2, \ldots, m$. With these choices, the operator Φ_m maps each $f \in \mathcal{H}$ to the m-vector of rescaled samples

$$\frac{1}{\sqrt{m}}[\, f(t_1) \quad \cdots \quad f(t_m)\,]^T.$$

  Defining the rescaled noise $\sigma_m = \frac{\sigma_0}{\sqrt{m}}$ yields an instantiation of model (10) which is equivalent to time-sampling (8).
- For the basis truncation model (9), we set $\phi_j = \mu_j \psi_j$ so that the operator Φ maps each function $f \in \mathcal{H}$ to the vector of basis coefficients $[\langle \psi_1, f\rangle_{L^2} \;\cdots\; \langle \psi_m, f\rangle_{L^2}]^T$. Setting $\sigma_m = \sigma_0$ then yields another instantiation of model (10), this one equivalent to basis truncation (9).

A remark on notation before proceeding: in the remainder of the paper, we use (Φ, σ) as shorthand notation for (Φ_m, σ_m), since the index m should be implicitly understood throughout our analysis.

In this paper, we provide and analyze estimators for the r-dimensional eigensubspace spanned by $\{f_j^*\}$, in both the sampled domain $\mathbb{R}^m$ and in the functional domain. To be more specific, for $j = 1, \ldots, r$, define the vectors $z_j^* := \Phi f_j^* \in \mathbb{R}^m$, and the subspaces

$$\mathfrak{Z}^* := \operatorname{span}\{z_1^*, \ldots, z_r^*\} \subset \mathbb{R}^m \quad \text{and} \quad \mathfrak{F}^* := \operatorname{span}\{f_1^*, \ldots, f_r^*\} \subset \mathcal{H}, \tag{11}$$

and let $\widehat{\mathfrak{Z}}$ and $\widehat{\mathfrak{F}}$ denote the corresponding estimators. In order to measure the performance of the estimators, we will use projection-based distances between subspaces. In particular, let $P_{\mathfrak{Z}^*}$ and $P_{\widehat{\mathfrak{Z}}}$ be orthogonal projection operators into $\mathfrak{Z}^*$ and $\widehat{\mathfrak{Z}}$, respectively, considered as subspaces of $\ell_2^m := (\mathbb{R}^m, \|\cdot\|_2)$. Similarly, let $P_{\mathfrak{F}^*}$ and $P_{\widehat{\mathfrak{F}}}$ be orthogonal projection operators into $\mathfrak{F}^*$ and $\widehat{\mathfrak{F}}$, respectively, considered as subspaces of $(\mathcal{H}, \|\cdot\|_{L^2})$. We are interested in bounding the deviations

$$\text{(12)} \qquad \mathrm{d_{HS}}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) := |||P_{\widehat{\mathfrak{Z}}} - P_{\mathfrak{Z}^*}|||_{\mathrm{HS}} \quad \text{and} \quad \mathrm{d_{HS}}(\widehat{\mathfrak{F}}, \mathfrak{F}^*) := |||P_{\widehat{\mathfrak{F}}} - P_{\mathfrak{F}^*}|||_{\mathrm{HS}},$$

where $|||\cdot|||_{\mathrm{HS}}$ is the Hilbert–Schmidt norm of an operator (or matrix).

2.3. *Approximation-theoretic quantities.* One object that plays an important role in our analysis is the matrix $K := \Phi\Phi^* \in \mathbb{R}^{m\times m}$. From the form of the adjoint, it can be seen that $[K]_{ij} = \langle \phi_i, \phi_j \rangle_{\mathcal{H}}$. For future reference, let us compute this matrix for the two special cases of linear operators considered thus far:

- For the time sampling model (8), we have $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$ for all $j = 1, \ldots, m$, and hence $[K]_{ij} = \frac{1}{m}\langle \mathbb{K}(\cdot, t_i), \mathbb{K}(\cdot, t_j)\rangle_{\mathcal{H}} = \frac{1}{m}\mathbb{K}(t_i, t_j)$, using the reproducing property of the kernel.
- For the basis truncation model (9), we have $\phi_j = \mu_j\psi_j$, and hence $[K]_{ij} = \langle \mu_i\psi_i, \mu_j\psi_j\rangle_{\mathcal{H}} = \mu_i\delta_{ij}$. Thus, in this special case, we have $K = \mathrm{diag}(\mu_1, \ldots, \mu_m)$.

In general, the matrix K is a type of Gram matrix, and so is symmetric and positive semidefinite. We assume throughout this paper that the functions $\{\phi_j\}_{j=1}^m$ are linearly independent in $\mathcal{H}$, which implies that K is strictly positive definite. Consequently, it has a set of eigenvalues which can be ordered as

$$\text{(13)} \qquad \widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots \geq \widehat{\mu}_m > 0.$$

Under this condition, we may use K to define a norm on $\mathbb{R}^m$ via $\|z\|_K^2 := z^T K^{-1} z$. Moreover, we have the following interpolation lemma, which is proved in Appendix B.1 of the supplementary material [3]:

LEMMA 1. *For any $f \in \mathcal{H}$, we have $\|\Phi f\|_K \leq \|f\|_{\mathcal{H}}$, with equality if and only if $f \in \mathrm{Ra}(\Phi^*)$. Moreover, for any $z \in \mathbb{R}^m$, the function $g = \Phi^* K^{-1} z$ has smallest Hilbert norm of all functions satisfying $\Phi g = z$, and is the unique function with this property.*

This lemma is useful in constructing a function-based estimator, as will be clarified in Section 3.

In our analysis of the functional error $\mathrm{d_{HS}}(\widehat{\mathfrak{F}}, \mathfrak{F}^*)$, a number of approximation-theoretic quantities play an important role. As a mapping from an infinite-dimensional space $\mathcal{H}$ to $\mathbb{R}^m$, the operator Φ has a nontrivial nullspace. Given the

observation model (10), we receive no information about any component of a function f^* that lies within this nullspace. For this reason, we define the width of the nullspace in the L^2-norm, namely the quantity

$$N_m(\Phi) := \sup\{\|f\|_{L^2}^2 \mid f \in \mathrm{Ker}(\Phi), \|f\|_{\mathcal{H}} \leq 1\}. \tag{14}$$

In addition, the observation operator Φ induces a semi-norm on the space $\mathcal{H}$, defined by

$$\|f\|_{\Phi}^2 := \|\Phi f\|_2^2 = \sum_{j=1}^{m} [\Phi f]_j^2. \tag{15}$$

It is of interest to assess how well this semi-norm approximates the L^2-norm. Accordingly, we define the quantity

$$D_m(\Phi) := \sup_{\substack{f \in \mathrm{Ra}(\Phi^*) \\ \|f\|_{\mathcal{H}} \leq 1}} |\|f\|_{\Phi}^2 - \|f\|_{L^2}^2|, \tag{16}$$

which measures the worst-case gap between these two (semi)-norms, uniformly over the Hilbert ball of radius one, restricted to the subspace of interest $\mathrm{Ra}(\Phi^*)$. Given knowledge of the linear operator Φ, the quantity $D_m(\Phi)$ can be computed in a relatively straightforward manner. In particular, recall the definition of the matrix K, and let us define a second matrix $\Theta \in \mathbb{S}_+^m$ with entries $\Theta_{ij} := \langle \varphi_i, \varphi_j \rangle_{L^2}$.

LEMMA 2. *We have the equivalence*

$$D_m(\Phi) = \left|\!\left|\!\left| K - K^{-1/2} \Theta K^{-1/2} \right|\!\right|\!\right|_2, \tag{17}$$

where $|\!|\!| \cdot |\!|\!|_2$ *denotes the* ℓ_2*-operator norm.*

See Appendix B.2 of the supplementary material [3] for the proof of this claim.

3. M-estimator and implementation. With this background in place, we now turn to the description of our M-estimator, as well as practical details associated with its implementation.

3.1. *M-estimator.* We begin with some preliminaries on notation, and our representation of subspaces. Recall definition (11) of $\mathfrak{Z}^*$ as the r-dimensional subspace of $\mathbb{R}^m$ spanned by $\{z_1^*, \ldots, z_r^*\}$, where $z_j^* = \Phi f_j^*$. Our initial goal is to construct an estimate $\widehat{\mathfrak{Z}}$, itself an r-dimensional subspace, of the unknown subspace $\mathfrak{Z}^*$.

We represent subspaces by elements of the Stiefel manifold $V_r(\mathbb{R}^m)$, which consists of $m \times r$ matrices Z with orthonormal columns

$$V_r(\mathbb{R}^m) := \{Z \in \mathbb{R}^{m \times r} \mid Z^T Z = I_r\}.$$

A given matrix Z acts as a representative of the subspace spanned by its columns, denoted by $\mathrm{col}(Z)$. For any $U \in V_r(\mathbb{R}^r)$, the matrix ZU also belongs to the Stiefel manifold, and since $\mathrm{col}(Z) = \mathrm{col}(ZU)$, we may call ZU a version of Z. We let $P_Z = ZZ^T \in \mathbb{R}^{m\times m}$ be the orthogonal projection onto $\mathrm{col}(Z)$. For two matrices $Z_1, Z_2 \in V_r(\mathbb{R}^m)$, we measure the distance between the associated subspaces via $\mathrm{d}_{\mathrm{HS}}(Z_1, Z_2) := |||P_{Z_1} - P_{Z_2}|||_{\mathrm{HS}}$, where $|||\cdot|||_{\mathrm{HS}}$ is the Hilbert–Schmidt (or Frobenius) matrix norm.

3.1.1. *Subspace-based estimator.* With this notation, we now specify an M-estimator for the subspace $\mathfrak{Z}^* = \mathrm{span}\{z_1^*, \ldots, z_r^*\}$. Let us begin with some intuition. Given the n samples $\{y_1, \ldots, y_n\}$, let us define the $m \times m$ sample covariance matrix $\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n y_i y_i^T$. Given the observation model (10), a straightforward computation shows that

$$\mathbb{E}[\widehat{\Sigma}_n] = \sum_{j=1}^r s_j^2 z_j^* (z_j^*)^T + \sigma_m^2 I_m. \tag{18}$$

Thus, as n becomes large, we expect that the top r eigenvectors of $\widehat{\Sigma}_n$ might give a good approximation to $\mathrm{span}\{z_1^*, \ldots, z_r^*\}$. By the Courant–Fischer variational representation, these r eigenvectors can be obtained by maximizing the objective function

$$\langle\langle \widehat{\Sigma}_n, P_Z \rangle\rangle := \mathrm{tr}\big(\widehat{\Sigma}_n ZZ^T\big)$$

over all matrices $Z \in V_r(\mathbb{R}^m)$.

However, this approach fails to take into account the smoothness constraints that the vectors $z_j^* = \Phi f_j^*$ inherit from the smoothness of the eigenfunctions f_j^*. Since $\|f_j^*\|_{\mathcal{H}} \le \rho$ by assumption, Lemma 1 implies that

$$\|z_j^*\|_K^2 = (z_j^*)^T K^{-1} z_j^* \le \|f_j^*\|_{\mathcal{H}}^2 \le \rho^2 \qquad \text{for all } j = 1, 2, \ldots, r.$$

Consequently, if we define the matrix $Z^* := [z_1^* \ \cdots \ z_r^*] \in \mathbb{R}^{m\times r}$, then it must satisfy the *trace smoothness condition*

$$\big\langle\big\langle K^{-1}, Z^*(Z^*)^T \big\rangle\big\rangle = \sum_{j=1}^r (z_j^*)^T K^{-1} z_j^* \le r\rho^2. \tag{19}$$

This calculation motivates the constraint $\langle\langle K^{-1}, P_Z\rangle\rangle \le 2r\rho^2$ in our estimation procedure.

Based on the preceding intuition, we are led to consider the optimization problem

$$\widehat{Z} \in \operatorname*{arg\,max}_{Z\in V_r(\mathbb{R}^m)} \big\{ \langle\langle \widehat{\Sigma}_n, P_Z\rangle\rangle \mid \big\langle\big\langle K^{-1}, P_Z\big\rangle\big\rangle \le 2r\rho^2 \big\}, \tag{20}$$

where we recall that $P_Z = ZZ^T \in \mathbb{R}^{m\times m}$. Given any optimal solution $\widehat{Z}$, we return the subspace $\widehat{\mathfrak{Z}} = \operatorname{col}(\widehat{Z})$ as our estimate of $\mathfrak{Z}^*$. As discussed at more length in Section 3.2, it is straightforward to compute $\widehat{Z}$ in polynomial time. The reader might wonder why we have included an additional factor of two in this trace smoothness condition. This slack is actually needed due to the potential infeasibility of the matrix Z^* for to problem (20), which arises since the columns of Z^* are not guaranteed to be orthonormal. As shown by our analysis, the additional slack allows us to find a matrix $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$ that spans the same subspace as Z^*, and is also feasible for to problem (20). More formally, we have:

LEMMA 3. *Under condition* (27b), *there exists a matrix* $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$ *such that*

$$\operatorname{Ra}(\widetilde{Z}^*) = \operatorname{Ra}(Z^*) \quad \text{and} \quad \langle\langle K^{-1}, \widetilde{Z}^*(\widetilde{Z}^*)^T\rangle\rangle \le 2r\rho^2. \tag{21}$$

See Appendix B.3 of the supplementary material [3] for the proof of this claim.

3.1.2. *The functional estimate* $\widehat{\mathfrak{F}}$. Having thus obtained an estimate[2] $\widehat{\mathfrak{Z}} = \operatorname{span}\{\widehat{z}_1, \ldots, \widehat{z}_r\}$ of $\mathfrak{Z}^* = \operatorname{span}\{z_1^*, \ldots, z_r^*\}$, we now need to construct a r-dimensional subspace $\widehat{\mathfrak{F}}$ of the Hilbert space to be used as an estimate of $\mathfrak{F}^* = \operatorname{span}\{f_1^*, \ldots, f_r^*\}$. We do so using the interpolation suggested by Lemma 1. For each $j = 1, \ldots, r$, let us define the function

$$\widehat{f}_j := \Phi^* K^{-1}\widehat{z}_j = \sum_{i=1}^{m} (K^{-1}\widehat{z}_j)_i \phi_i. \tag{22}$$

Since $K = \Phi\Phi^*$ by definition, this construction ensures that $\Phi\widehat{f}_j = \widehat{z}_j$. Moreover, Lemma 1 guarantees that $\widehat{f}_j$ has the minimal Hilbert norm (and hence is smoothest in a certain sense) over all functions that have this property. Finally, since Φ is assumed to be surjective (equivalently, K assumed invertible), $\Phi^* K^{-1}$ maps linearly independent vectors to linearly independent functions, and hence preserves dimension. Consequently, the space $\widehat{\mathfrak{F}} := \operatorname{span}\{\widehat{f}_1, \ldots, \widehat{f}_r\}$ is an r-dimensional subspace of $\mathcal{H}$ that we take as our estimate of $\mathfrak{F}^*$.

3.2. *Implementation details*. In this section, we consider some practical aspects of implementing the M-estimator, and present some simulations to illustrate its qualitative properties. We begin by observing that once the subspace vectors $\{\widehat{z}_j\}_{j=1}^r$ have been computed, then it is straightforward to compute the function estimates $\{\widehat{f}_j\}_{j=1}^r$, as weighted combinations of the functions $\{\phi_j\}_{j=1}^m$. Accordingly, we focus our attention on solving problem (20).

[2]Here, $\{\widehat{z}_j\}_{j=1}^r \subset \mathbb{R}^m$ is any collection of vectors that span $\widehat{\mathfrak{Z}}$. As we are ultimately only interested in the resulting functional "subspace," it does not matter which particular collection we choose.

On the surface, problem (20) might appear nonconvex, due to the Stiefel manifold constraint. However, it can be reformulated as a semidefinite program (SDP), a well-known class of convex programs, as clarified in the following:

LEMMA 4. *Problem* (20) *is equivalent to solving the SDP*

$$\widehat{X} \in \arg\max_{X \succeq 0} \langle\langle \widehat{\Sigma}_n, X \rangle\rangle \quad \text{such that } |||X|||_2 \le 1, \operatorname{tr}(X) = r, \text{ and } \langle\langle K^{-1}, X \rangle\rangle \le 2r\rho^2 \tag{23}$$

for which there always exists an optimal rank r solution. Moreover, by Lagrangian duality, for some $\beta > 0$, *the problem is equivalent to*

$$\widehat{X} \in \arg\max_{X \succeq 0} \langle\langle \widehat{\Sigma}_n - \beta K^{-1}, X \rangle\rangle \quad \text{such that } |||X|||_2 \le 1 \text{ and } \operatorname{tr}(X) = r, \tag{24}$$

which can be solved by an eigen decomposition of $\widehat{\Sigma}_n - \beta K^{-1}$.

As a consequence, for a given Lagrange multiplier β, the regularized form of the estimator can be solved with the cost of solving an eigenvalue problem. For a given constraint $2r\rho^2$, the appropriate value of β can be found by a path-tracing algorithm, or a simple dyadic splitting approach.

In practice where the radius ρ is not known, one could use cross-validation to set a proper value for the Lagrange multiplier β. A possibly simpler approach is to evaluate $\langle\langle K^{-1}, X \rangle\rangle$ for the optimal X on a grid of β and choose a value around which $\langle\langle K^{-1}, X \rangle\rangle$ is least variable. As for the choice of the number of components r, a standard approach for choosing it would be to compute the estimator for different choices, and plot the residual sum of eigenvalues of the sample covariance matrix. As in ordinary PCA, an elbow in such a plot indicates a proper trade-off between the number of components to keep and the amount of variation explained.

In order to illustrate the estimator, we consider the time sampling model (8), with uniformly spaced samples, in the context of a first-order Sobolev RKHS [with kernel function $\mathbb{K}(s,t) = \min(s,t)$]. The parameters of the model are taken to be $r = 4$, $(s_1, s_2, s_3, s_4) = (1, 0.5, 0.25, 0.125)$, $\sigma_0 = 1$, $m = 100$ and $n = 75$. The regularized form (24) of the estimator is applied, and the results are shown in Figure 1. The top row corresponds to the four "true" signals $\{f_j^*\}$, the leftmost being f_1^* (i.e., having the highest signal-to-noise ratio) and the rightmost f_4^*. The subsequent rows show the corresponding estimates $\{\widehat{f}_j\}$, obtained using different values of β. The second, third and fourth rows correspond to $\beta = 0$, $\beta = 0.0052$ and $\beta = 0.83$.

One observes that without regularization ($\beta = 0$), the estimates for the two weakest signals (f_3^* and f_4^*) are poor. The case $\beta = 0.0052$ is roughly the one which achieves the minimum for the dual problem. One observes that the quality

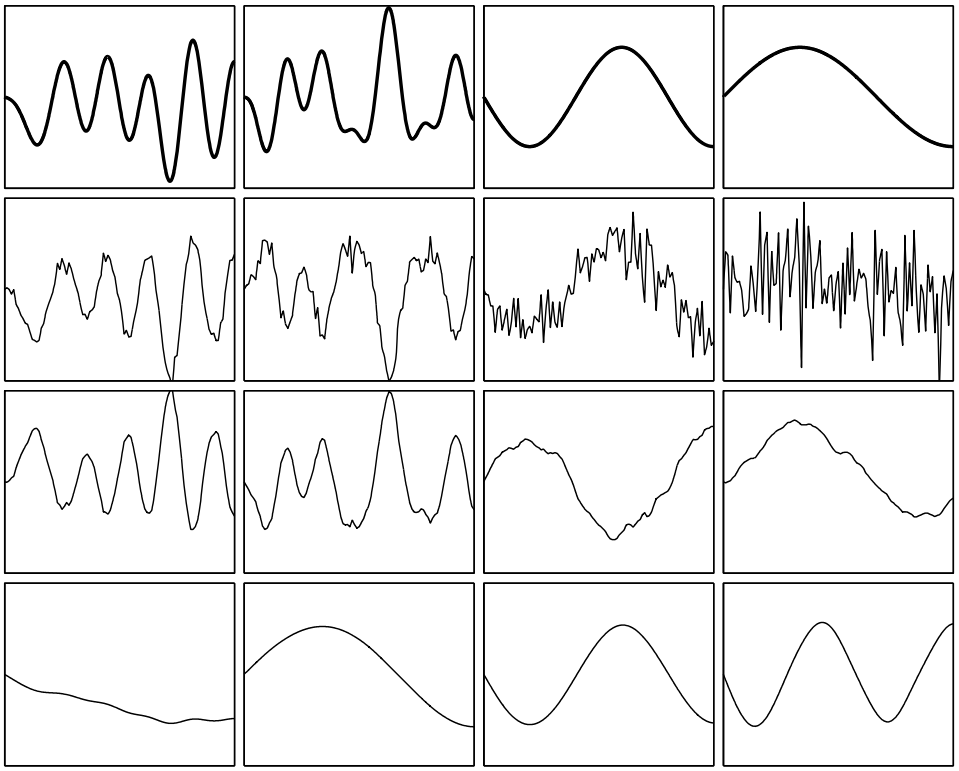

FIG. 1. *Regularized PCA for time sampling in first-order Sobolev RKHS. Top row shows, from left to right, plots of the* $r = 4$ *"true" principal components* $f_1^*, \ldots, f_4^*$ *with signal-to-noise ratios* $s_1 = 1, s_2 = 0.5, s_3 = 0.25$ *and* $s_4 = 0.125$, *respectively. The number of statistical and functional samples are* $n = 75$ *and* $m = 100$. *Subsequent rows show the corresponding estimators* $\widehat{f}_1, \ldots, \widehat{f}_4$ *obtained by applying the regularized form* (24).

of the estimates of the signals, and in particular the weakest ones, are considerably improved. The optimal (oracle) value of β, that is, the one which achieves the minimum error between $\{f_j^*\}$ and $\{\widehat{f}_j\}$, is $\beta = 0.0075$ in this problem. The corresponding estimates are qualitatively similar to those of $\beta = 0.0052$ and are not shown.

The case $\beta = 0.83$ shows the effect of over-regularization. It produces very smooth signals, and although it fails to reveal f_1^* and f_2^*, it reveals highly accurate versions of f_3^* and f_4^*. It is also interesting to note that the smoothest signal, f_4^*, now occupies the position of the second (estimated) principal component. That is, the regularized PCA sees an effective signal-to-noise ratio which is influenced by smoothness. This suggests a rather practical appeal of the method in revealing smooth signals embedded in noise. One can vary β from zero upward, and if some patterns seem to be present for a wide range of β (and getting smoother as β is

increased), one might suspect that they are indeed present in data but masked by noise.

4. Main results. We now turn to the statistical analysis of our estimators, in particular deriving high-probability upper bounds on the error of the subspace-based estimate $\widehat{\mathfrak{Z}}$, and the functional estimate $\widehat{\mathfrak{F}}$. In both cases, we begin by stating general theorems that apply to arbitrary linear operators Φ—Theorems 1 and 2, respectively—and then derive a number of corollaries for particular instantiations of the observation operator.

4.1. *Subspace-based estimation rates (for $\widehat{\mathfrak{Z}}$).* We begin by stating high-probability upper bounds on the error $\mathrm{d}_{\mathrm{HS}}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*)$ of the subspace-based estimates. Our rates are stated in terms of a function that involves the eigenvalues of the matrix $K = \Phi\Phi^* \in \mathbb{R}^m$, ordered as $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots \geq \widehat{\mu}_m > 0$. Consider the function $\mathcal{F}: \mathbb{R}_+ \to \mathbb{R}_+$ given by

$$\mathcal{F}(t) := \left[\sum_{j=1}^{m} \min\{t^2, r\rho^2\widehat{\mu}_j\}\right]^{1/2}. \tag{25}$$

As will be clarified in our proofs, this function provides a measure of the statistical complexity of the function class

$$\mathrm{Ra}(\Phi^*) = \left\{ f \in \mathcal{H} \;\middle|\; f = \sum_{j=1}^{m} a_j \phi_j \text{ for some } a \in \mathbb{R}^m \right\}.$$

We require a few regularity assumptions. Define the quantity

$$C_m(f^*) := \max_{1 \leq i,j \leq r} \left|\langle f_i^*, f_j^* \rangle_\Phi - \delta_{ij}\right| = \max_{1 \leq i,j \leq r} \left|\langle z_i^*, z_j^* \rangle_{\mathbb{R}^m} - \delta_{ij}\right|, \tag{26}$$

which measures the departure from orthonormality of the vectors $z_j^* := \Phi f_j^*$ in $\mathbb{R}^m$. A straightforward argument using a polarization identity shows that $C_m(f^*)$ is upper bounded (up to a constant factor) by the uniform quantity $D_m(\Phi)$, as defined in equation (16). Recall that the random functions are generated according to the model $x_i = \sum_{j=1}^{r} s_j \beta_{ij} f_j^*$, where the signal strengths are ordered as $1 = s_1 \geq s_2 \geq \cdots \geq s_r > 0$, and that σ_m denotes the noise standard deviation in the observation model (10).

In terms of these quantities, we require the following assumptions:

$$\text{(A1)} \quad \frac{s_r^2}{s_1^2} \geq \frac{1}{2} \quad \text{and} \quad \sigma_0^2 := \sup_m \sigma_m^2 \leq \kappa s_1^2, \tag{27a}$$

$$\text{(A2)} \quad C_m(f^*) \leq \frac{1}{2r} \quad \text{and} \tag{27b}$$

$$\text{(A3)} \quad \frac{\sigma_m}{\sqrt{n}} \mathcal{F}(t) \leq \sqrt{\kappa} t \qquad \text{for the same constant } \kappa \text{ as in (A1)}, \tag{27c}$$

$$\text{(A4)} \quad r \leq \min\left\{\frac{m}{2}, \frac{n}{4}, \kappa\frac{\sqrt{n}}{\sigma_m}\right\}. \tag{27d}$$

REMARKS. The first part of condition (A1) is to prevent the ratio s_r/s_1 from going to zero as the pair (m, n) increases, where the constant $1/2$ is chosen for convenience. Such a lower bound is necessary for consistent estimation of the eigensubspace corresponding to $\{s_1, \ldots, s_r\}$. The second part of condition (A1), involving the constant κ, provides a lower bound on the signal-to-noise ratio s_r/σ_m. Condition (A2) is required to prevent degeneracy among the vectors $z_j^* = \Phi f_j^*$ obtained by mapping the unknown eigenfunctions to the observation space $\mathbb{R}^m$. [In the ideal setting, we would have $C_m(f^*) = 0$, but our analysis shows that the upper bound in (A2) is sufficient.] Condition (A3) is required so that the critical tolerance $\epsilon_{m,n}$ specified below is well-defined; as will be clarified, it is always satisfied for the time-sampling model, and holds for the basis truncation model whenever $n \geq m$. Condition (A4) is easily satisfied, since the right-hand side of (27d) goes to infinity while we usually take r to be fixed. Our results, however, are still valid if r grows slowly with m and n subject to (27d).

THEOREM 1. *Under conditions* (A1)–(A3) *for a sufficiently small constant* κ, *let* $\epsilon_{m,n}$ *be the smallest positive number satisfying the inequality*

$$\frac{\sigma_m}{\sqrt{n}} r^{3/2} \mathcal{F}(\epsilon) \leq \kappa \epsilon^2. \tag{28}$$

Then there are universal positive constants (c_0, c_1, c_2) *such that*

$$\mathbb{P}\big[\mathrm{d}_{\mathrm{HS}}^2(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \leq c_0 \epsilon_{m,n}^2\big] \geq 1 - \varphi(n, \epsilon_{m,n}), \tag{29}$$

where $\varphi(n, \epsilon_{m,n}) := c_1\{r^2 \exp(-c_2 r^{-3} \frac{n}{\sigma_m^2}(\epsilon_{m,n} \wedge \epsilon_{m,n}^2)) + r\exp(-\frac{n}{64})\}$.

We note that Theorem 1 is a general result, applying to an arbitrary bounded linear operator Φ. However, we can obtain a number of concrete results by making specific choices of this sampling operator, as we explore in the following sections.

4.1.1. *Consequences for time-sampling.* Let us begin with the time-sampling model (8), in which we observe the sampled functions

$$y_i = [\, x_i(t_1) \quad x_i(t_2) \quad \cdots \quad x_i(t_m)\,]^T + \sigma_0 w_i \qquad \text{for } i = 1, 2, \ldots, m.$$

As noted earlier, this set-up can be modeled in our general setting (10) with $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$ and $\sigma_m = \sigma_0/\sqrt{m}$.

In this case, by the reproducing property of the RKHS, the matrix $K = \Phi\Phi^*$ has entries of the form $K_{ij} = \langle \phi_i, \phi_j \rangle_{\mathcal{H}} = \frac{\mathbb{K}(t_i, t_j)}{m}$. Letting $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \cdots \geq \widehat{\mu}_m > 0$ denote its ordered eigenvalues, we say that the kernel matrix K has polynomial-decay with parameter $\alpha > 1/2$ if there is a constant c such that $\widehat{\mu}_j \leq cj^{-2\alpha}$ for all

$j = 1, 2, \ldots, m$. Since the kernel matrix K represents a discretized approximation of the kernel integral operator defined by $\mathbb{K}$, this type of polynomial decay is to be expected whenever the kernel operator has polynomial-α decaying eigenvalues. For example, the usual spline kernels that define Sobolev spaces have this type of polynomial decay [14]. In Appendix A of the supplementary material [3], we verify this property explicitly for the kernel $\mathbb{K}(s,t) = \min\{s,t\}$ that defines the Sobolev class with smoothness $\alpha = 1$.

For any such kernel, we have the following consequence of Theorem 1:

COROLLARY 1 (Achievable rates for time-sampling). *Consider the case of a time-sampling operator* Φ. *In addition to conditions* (A1) *and* (A2), *suppose that the kernel matrix* K *has polynomial-decay with parameter* $\alpha > 1/2$. *Then we have*

$$\mathbb{P}\left[\mathrm{d}_{\mathrm{HS}}^2(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \le c_0 \min\left\{\left(\frac{\kappa_{r,\rho}\sigma_0^2}{mn}\right)^{2\alpha/(2\alpha+1)}, r^3\frac{\sigma_0^2}{n}\right\}\right] \ge 1 - \varphi(n,m), \tag{30}$$

where $\kappa_{r,\rho} := r^{3+1/(2\alpha)}\rho^{1/\alpha}$, *and* $\varphi(n,m) := c_1\{\exp(-c_2\{(r^{-2}\rho^2 mn)^{1/(2\alpha+1)} \wedge m\}) + \exp(-n/64)\}$.

REMARKS. (a) Disregarding constant pre-factors not depending on the pair (m,n), Corollary 1 guarantees that solving problem (20) returns a subspace estimate $\widehat{\mathfrak{Z}}$ such that

$$\mathrm{d}_{\mathrm{HS}}^2(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \precsim \min\{(mn)^{-2\alpha/(2\alpha+1)}, n^{-1}\}$$

with high probability as (m,n) increase. Depending on the scaling of the number of time samples m relative to the number of functional samples n, either term in this upper bound can be the smallest (and hence active) one. For instance, it can be verified that whenever $m \ge n^{1/(2\alpha)}$, then the first term is smallest, so that we achieve the rate $\mathrm{d}_{\mathrm{HS}}^2(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \precsim (mn)^{-2\alpha/(2\alpha+1)}$. The appearance of the term $(mn)^{-2\alpha/(2\alpha+1)}$ is quite natural, as it corresponds to the minimax rate of a nonparametric regression problem with smoothness α, based on m samples each of variance n^{-1}. Later, in Section 4.3, we provide results guaranteeing that this scaling is minimax optimal under reasonable conditions on the choice of sample points; in particular, see Theorem 3(a).

(b) To be clear, although bound (30) allows for the possibility that the error is of order *lower than* n^{-1}, we note that the probability with which the guarantee holds includes a term of the order $\exp(-n/64)$. Consequently, in terms of expected error, we cannot guarantee a rate faster than n^{-1}.

PROOF OF COROLLARY 1. We need to bound the critical value $\epsilon_{m,n}$ defined in the theorem statement (28). Define the function $\mathcal{G}^2(t) := \sum_{j=1}^m \min\{\widehat{\mu}_j, t^2\}$, and note that $\mathcal{F}(t) = \sqrt{r}\rho\mathcal{G}(\frac{t}{\sqrt{r}\rho})$ by construction. Under the assumption of polynomial-α eigendecay, we have

$$\mathcal{G}^2(t) \le \int_0^\infty \min\{cx^{-2\alpha}, t^2\}\,dx,$$

and some algebra then shows that $\mathcal{G}(t) \precsim t^{1-1/(2\alpha)}$. Disregarding constant factors, an upper bound on the critical $\epsilon_{m,n}$ can be obtained by solving the equation

$$\epsilon^2 = \frac{\sigma_m}{\sqrt{n}} r^{3/2} \sqrt{r}\rho \left(\frac{\epsilon}{\sqrt{r}\rho} \right)^{1-1/(2\alpha)}.$$

Doing so yields the upper bound $\epsilon^2 \precsim [\frac{\sigma_m^2}{n} r^3 (\sqrt{r}\rho)^{1/\alpha}]^{2\alpha/(2\alpha+1)}$. Otherwise, we also have the trivial upper bound $\mathcal{F}(t) \leq \sqrt{m}t$, which yields the alternative upper bound $\varepsilon_{m,n} \precsim (\frac{m\sigma_m^2}{n} r^3)^{1/2}$. Recalling that $\sigma_m = \sigma_0/\sqrt{m}$ and combining the pieces yields the claim. Notice that this last (trivial) bound on $\mathcal{F}(t)$ implies that condition (A3) is always satisfied for the time-sampling model. □

4.1.2. *Consequences for basis truncation.* We now turn to some consequences for the basis truncation model (9).

COROLLARY 2 (Achievable rates for basis truncation). *Consider a basis truncation operator Φ in a Hilbert space with polynomial-α decay. Under conditions* (A1), (A2) *and $m \leq n$, we have*

$$\mathbb{P}\left[\mathrm{d}_{\mathrm{HS}}^2(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*) \leq c_0 \left(\frac{\kappa_{r,\rho}\sigma_0^2}{n} \right)^{2\alpha/(2\alpha+1)} \right] \geq 1 - \varphi(n,m), \tag{31}$$

where $\kappa_{r,\rho} := r^{3+1/(2\alpha)}\rho^{1/\alpha}$, and $\varphi(n,m) := c_1\{\exp(-c_2(r^{-2}\rho^2 n)^{1/(2\alpha+1)}) + \exp(-n/64)\}$.

PROOF. We note that as long as $m \leq n$, condition (A3) is satisfied, since $\frac{\sigma_m}{\sqrt{n}}\mathcal{F}(t) \leq \sigma_0\sqrt{\frac{m}{n}}t \leq \sigma_0 t$. The rest of the proof follows that of Corollary 1, noting that in the last step we have $\sigma_m = \sigma_0$ for the basis truncation model. □

4.2. *Function-based estimation rates (for $\widehat{\mathfrak{F}}$).* As mentioned earlier, given the consistency of $\widehat{\mathfrak{Z}}$, the consistency of $\widehat{\mathfrak{F}}$ is closely related to approximation properties of the semi-norm $\|\cdot\|_\Phi$ induced by Φ, and in particular how closely it approximates the L^2-norm. These approximation-theoretic properties are captured in part by the nullspace width $N_m(\Phi)$ and defect $D_m(\Phi)$ defined earlier in equations (14) and (16), respectively. In addition to these previously defined quantities, we require bounds on the following global quantity:

$$R_m(\epsilon; \nu) := \sup\{\|f\|_{L^2}^2 \mid \|f\|_{\mathcal{H}}^2 \leq \nu^2, \|f\|_\Phi^2 \leq \epsilon^2\}. \tag{32}$$

A general upper bound on this quantity is of the form

$$R_m(\epsilon; \nu) \leq c_1\epsilon^2 + \nu^2 S_m(\Phi). \tag{33}$$

In fact, it is not hard to show that such a bound exists with $c_1 = 2$ and $S_m(\Phi) = 2(D_m(\Phi) + N_m(\Phi))$ using the decomposition $\mathcal{H} = \mathrm{Ra}(\Phi^*) \oplus \mathrm{Ker}(\Phi)$.

However, this bound is not sharp. Instead, one can show that in most cases of interest, the term $S_m(\Phi)$ is of the order of $N_m(\Phi)$.

There are a variety of conditions that ensure that $S_m(\Phi)$ has this scaling; we refer the reader to the paper [2] for a general approach. Here we provide a simple sufficient condition, namely,

$$\text{(B1)} \quad \Theta \preceq c_0 K^2 \tag{34}$$

for a positive constant c_0.

LEMMA 5. *Under* (B1), *bound* (33) *holds with* $c_1 = 2c_0$ *and* $S_m(\Phi) = 2N_m(\Phi)$.

See Appendix B.4 of the supplementary material [3] for the proof of this claim. In the sequel, we show that the first-order Sobolev RKHS satisfies condition (B1).

THEOREM 2. *Suppose that condition* (A1) *holds, and the approximation-theoretic quantities satisfy the bounds* $D_m(\Phi) \leq \frac{1}{4r\rho^2} \leq 1$ *and* $N_m(\Phi) \leq 1$. *Then there is a constant* $\kappa'_{r,\rho}$ *such that*

$$\mathrm{d}^2_{\mathrm{HS}}(\widehat{\mathfrak{F}}, \mathfrak{F}^*) \leq \kappa'_{r,\rho}\{\epsilon^2_{m,n} + S_m(\Phi) + [D_m(\Phi)]^2\} \tag{35}$$

with the same probability as in Theorem 1.

As with Theorem 1, this is a generally applicable result, stated in abstract form. By specializing it to different sampling models, we can obtain concrete rates, as illustrated in the following sections.

4.2.1. *Consequences for time-sampling.* We begin by returning to the case of the time sampling model (8), where $\phi_j = \mathbb{K}(\cdot, t_j)/\sqrt{m}$. In this case, condition (B1) needs to be verified by some calculations. For instance, as shown in Appendix A of the supplementary material [3], in the case of the Sobolev kernel with smoothness $\alpha = 1$ [namely, $\mathbb{K}(s,t) = \min\{s,t\}$], we are guaranteed that (B1) holds with $c_0 = 1$, whenever the samples $\{t_j\}$ are chosen uniformly over $[0,1]$; hence, by Lemma 5, $S_m(\Phi) = 2N_m(\Phi)$. Moreover, in the case of uniform sampling, we expect that the nullspace width $N_m(\Phi)$ is upper bounded by μ_{m+1}, and so will be proportional to $m^{-2\alpha}$ in the case of a kernel operator with polynomial-α decay. This is verified in [2] (up to a logarithmic factor) for the case of the first-order Sobolev kernel. In Appendix A of the supplementary material [3], we also show that, for this kernel, $[D_m(\Phi)]^2$ is of the order $m^{-2\alpha}$, that is, of the same order as $N_m(\Phi)$.

COROLLARY 3. *Consider the basis truncation model* (9) *with uniformly spaced samples, and assume condition* (B1) *holds and that* $N_m(\Phi) + [D_m(\Phi)]^2 \precsim m^{-2\alpha}$. *Then the M-estimator returns a subspace estimate* $\widehat{\mathfrak{F}}$ *such that*

$$\mathrm{d}^2_{\mathrm{HS}}(\widehat{\mathfrak{F}}, \mathfrak{F}^*) \leq \kappa'_{r,\rho}\left\{\min\left\{\left(\frac{\sigma_0^2}{nm}\right)^{2\alpha/(2\alpha+1)}, \frac{\sigma_0^2}{n}\right\} + \frac{1}{m^{2\alpha}}\right\} \tag{36}$$

with the same probability as in Corollary 1.

In this case, there is an interesting trade-off between the bias or approximation error which is of order $m^{-2\alpha}$ and the estimation error. An interesting transition occurs at the point when $m \gtrsim n^{1/(2\alpha)}$, at which:

- the bias term $m^{-2\alpha}$ becomes of the order n^{-1}, so that it is no longer dominant, and
- for the two terms in the estimation error, we have the ordering

$$(mn)^{-2\alpha/(2\alpha+1)} \leq \left(n^{1+1/(2\alpha)}\right)^{-2\alpha/(2\alpha+1)} = n^{-1}.$$

Consequently, we conclude that the scaling $m = n^{1/(2\alpha)}$ is the minimal number of samples such that we achieve an overall bound of the order n^{-1} in the time-sampling model. In Section 4.3, we will see that these rates are minimax-optimal.

4.2.2. *Consequences for basis truncation.* For the basis truncation operator Φ, we have $\Theta = K^2 = \operatorname{diag}(\mu_1^2, \ldots, \mu_m^2)$ so that condition (B1) is satisfied trivially with $c_0 = 1$. Moreover, Lemma 2 implies $D_m(\Phi) = 0$. In addition, a function $f = \sum_{j=1}^{\infty} \sqrt{\mu_j} a_j \psi_j$ satisfies $\Phi f = 0$ if and only if $a_1 = a_2 = \cdots = a_m = 0$, so that

$$N_m(\Phi) = \sup\{\|f\|_{L^2}^2 \mid \|f\|_{\mathcal{H}} \leq 1, \Phi f = 0\} = \mu_{m+1}.$$

Consequently, we obtain the following corollary of Theorem 2:

COROLLARY 4. *Consider the basis truncation model (9) with a kernel operator that has polynomial-α decaying eigenvalues. Then the M-estimator returns a function subspace estimate $\widehat{\mathfrak{F}}$ such that*

$$\mathrm{d}_{\mathrm{HS}}^2(\widehat{\mathfrak{F}}, \mathfrak{F}^*) \leq \kappa'_{r,\rho} \left\{ \left(\frac{\sigma_0^2}{n} \right)^{2\alpha/(2\alpha+1)} + \frac{1}{m^{2\alpha}} \right\} \tag{37}$$

with the same probability as in Corollary 2.

By comparison to Corollary 3, we see that the trade-offs between (m, n) are very different for basis truncation. In particular, there is *no interaction* between the number of functional samples m and the number of statistical samples n. Increasing m only reduces the approximation error, whereas increasing n only reduces the estimation error. Moreover, in contrast to the time sampling model of Corollary 3, it is impossible to achieve the fast rate n^{-1}, regardless of how we choose the pair (m, n). In Section 4.3, we will also see that the rates given in Corollary 4 are minimax optimal.

4.3. *Lower bounds*. We now turn to lower bounds on the minimax risk, demonstrating the sharpness of our achievable results in terms of their scaling with (m, n). In order to do so, it suffices to consider the simple model with a single functional component $f^* \in \mathbb{B}_{\mathcal{H}}(1)$, so that we observe $y_i = \beta_{i1}\Phi_m(f^*) + \sigma_m w_i$ for $i = 1, 2, \ldots, n$, where $\beta_{i1} \sim N(0, 1)$ are i.i.d. standard normal variates. The minimax risk over the unit ball of the function space $\mathcal{H}$ in the Φ-norm is given by

$$\mathcal{M}_{m,n}^{\mathcal{H}}(\|\cdot\|_\Phi) := \inf_{\widetilde{f}} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}(1)} \mathbb{E}\|\widetilde{f} - f^*\|_\Phi^2, \tag{38}$$

where the function f^* ranges over the unit ball $\mathbb{B}_{\mathcal{H}}(1) = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$ of some Hilbert space, and $\widetilde{f}$ ranges over measurable functions of the data matrix $(y_1, y_2, \ldots, y_n) \in \mathbb{R}^{m \times n}$.

THEOREM 3 (Lower bounds for $\|\widetilde{f} - f^*\|_\Phi$). *Suppose that the kernel matrix K has eigenvalues with polynomial-α decay and* (A1) *holds.*

(a) *For the time-sampling model, we have*

$$\mathcal{M}_{m,n}^{\mathcal{H}}(\|\cdot\|_\Phi) \geq C \min\left\{\left(\frac{\sigma_0^2}{mn}\right)^{2\alpha/(2\alpha+1)}, \frac{\sigma_0^2}{n}\right\}. \tag{39}$$

(b) *For the frequency-truncation model, with $m \geq (c_0 n)^{1/(2\alpha+1)}$, we have*

$$\mathcal{M}_{m,n}^{\mathcal{H}}(\|\cdot\|_\Phi) \geq C\left(\frac{\sigma_0^2}{n}\right)^{2\alpha/(2\alpha+1)}. \tag{40}$$

Note that part (a) of Theorem 3 shows that the rates obtained in Corollary 3 for the case of time-sampling are minimax optimal. Similarly, comparing part (b) of the theorem to Corollary 4, we conclude that the rates obtained for frequency truncation model are minimax optimal for $n \in [m, c_1 m^{2\alpha+1}]$. As will become clear momentarily (as a consequence of our next theorem), the case $n > c_1 m^{2\alpha+1}$ is not of practical interest.

We now turn to lower bounds on the minimax risk in the $\|\cdot\|_{L^2}$ norm—namely

$$\mathcal{M}_{m,n}^{\mathcal{H}}(\|\cdot\|_{L^2}) := \inf_{\widetilde{f}} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}(1)} \mathbb{E}\|\widetilde{f} - f^*\|_{L^2}^2. \tag{41}$$

Obtaining lower bounds on this minimax risk requires another approximation property of the norm $\|\cdot\|_\Phi$ relative to $\|\cdot\|_{L^2}$. Consider matrix $\Psi \in \mathbb{R}^{m \times m}$ with entries $\Psi_{ij} := \langle \psi_i, \psi_j \rangle_\Phi$. Since the eigenfunctions are orthogonal in L^2, the deviation of Ψ from the identity measures how well the inner product defined by Φ approximates the L^2-inner product over the first m eigenfunctions of the kernel operator. For proving lower bounds, we require an upper bound of the form

$$\lambda_{\max}(\Psi) \leq c_1 \tag{B2}$$

for some universal constant $c_1 > 0$. As the proof will clarify, this upper bound is necessary in order that the Kullback–Leibler divergence—which controls the relative discriminability between different models—can be upper bounded in terms of the L^2-norm.

THEOREM 4 (Lower bounds for $\|\widetilde{f} - f^*\|_{L^2}^2$). *Suppose that condition* (B2) *holds, and the operator associated with kernel function $\mathbb{K}$ of the reproducing kernel Hilbert space $\mathcal{H}$ has eigenvalues with polynomial-α decay.*

(a) *For the time-sampling model, the minimax risk is lower bounded as*

$$\mathcal{M}_{m,n}^{\mathcal{H}}(\|\cdot\|_{L^2}) \geq C\left\{\min\left\{\left(\frac{\sigma_0^2}{mn}\right)^{2\alpha/(2\alpha+1)}, \frac{\sigma_0^2}{n}\right\} + \left(\frac{1}{m}\right)^{2\alpha}\right\}. \tag{42}$$

(b) *For the frequency-truncation model, the minimax error is lower bounded as*

$$\mathcal{M}_{m,n}^{\mathcal{H}}(\|\cdot\|_{L^2}) \geq C\left\{\left(\frac{\sigma_0^2}{n}\right)^{2\alpha/(2\alpha+1)} + \left(\frac{1}{m}\right)^{2\alpha}\right\}. \tag{43}$$

Verifying condition (B2) requires, in general, some calculations in the case of the time-sampling model. It is verified for uniform time-sampling for the first-order Sobolev RKHS in Appendix A of the supplementary material [3]. For the frequency-truncation model, condition (B2) always holds trivially since $\Psi = I_m$. By this theorem, the L^2 convergence rates of Corollaries 3 and 4 are minimax optimal. Also note that due to the presence of the approximation term $m^{-2\alpha}$ in (43), the Φ-norm term $n^{2\alpha/(2\alpha+1)}$ is only dominant when $m \geq c_2 n^{1/(2\alpha+1)}$ implying that this is the interesting regime for Theorem 3(b).

5. Proof of subspace-based rates. We now turn to the proofs of the results involving the error $\mathrm{d_{HS}}(\widehat{\mathfrak{Z}}, \mathfrak{Z}^*)$ between the estimated $\widehat{\mathfrak{Z}}$ and true subspace $\mathfrak{Z}^*$. We begin by proving Theorem 1, and then turn to its corollaries.

5.1. *Preliminaries.* We begin with some preliminaries before proceeding to the heart of the proof. Let us first introduce some convenient notation. Consider the $n \times m$ matrices

$$Y := [\, y_1 \quad y_2 \quad \cdots \quad y_n \,]^T \quad \text{and} \quad W := [\, w_1 \quad w_2 \quad \cdots \quad w_n \,]^T,$$

corresponding to the observation matrix Y and noise matrix W, respectively. In addition, we define the matrix $B := (\beta_{ij}) \in \mathbb{R}^{n \times r}$, and the diagonal matrix $S := \mathrm{diag}(s_1, \ldots, s_r) \in \mathbb{R}^{r \times r}$. Recalling that $Z^* := (z_1^*, \ldots, z_r^*) \in \mathbb{R}^{m \times r}$, the observation model (10) can be written in the matrix form $Y = B(Z^*S)^T + \sigma_m W$. Moreover, let us define the matrices $\overline{B} := \frac{B^T B}{n} \in \mathbb{R}^{r \times r}$ and $\overline{W} := \frac{W^T B}{n} \in \mathbb{R}^{m \times r}$.

Using this notation, some algebra shows that the associated sample covariance $\widehat{\Sigma}_n := \frac{1}{n} Y^T Y$ can be written in the form

$$\widehat{\Sigma}_n = \underbrace{Z^* S \overline{B} S (Z^*)^T}_{\Gamma} + \Delta_1 + \Delta_2, \tag{44}$$

where $\Delta_1 := \sigma_m [\overline{W} S (Z^*)^T + Z^* S \overline{W}^T]$ and $\Delta_2 := \sigma_m^2 \frac{W^T W}{n}$.

Lemma 3, proved in Appendix B.3 of the supplementary material [3], establishes the existence of a matrix $\widetilde{Z}^* \in V_r(\mathbb{R}^m)$ such that $\mathrm{Ra}(\widetilde{Z}^*) = \mathrm{Ra}(Z^*)$. As discussed earlier, due to the nature of the Steifel manifold, there are many versions of this matrix $\widetilde{Z}^*$, and also of any optimal solution matrix $\widehat{Z}$, obtained via right multiplication with an orthogonal matrix. For the subsequent arguments, we need to work with a particular version of $\widetilde{Z}^*$ (and $\widehat{Z}$) that we describe here.

Let us fix some convenient versions of $\widetilde{Z}^*$ and $\widehat{Z}$. As a consequence of CS decomposition, as long as $r \le m/2$, there exist orthogonal matrices $U, V \in \mathbb{R}^{r \times r}$ and an orthogonal matrix $Q \in \mathbb{R}^{m \times m}$ such that

$$Q^T \widetilde{Z}^* U = \begin{pmatrix} I_r \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad Q^T \widehat{Z} V = \begin{pmatrix} \widehat{C} \\ \widehat{S} \\ 0 \end{pmatrix}, \tag{45}$$

where $\widehat{C} = \mathrm{diag}(\widehat{c}_1, \ldots, \widehat{c}_r)$ and $\widehat{S} = \mathrm{diag}(\widehat{s}_1, \ldots, \widehat{s}_r)$ such that $1 \ge \widehat{s}_1 \ge \cdots \ge \widehat{s}_r \ge 0$ and $\widehat{C}^2 + \widehat{S}^2 = I_r$. See Bhatia [6], Theorem VII.1.8, for details on this decomposition. In the analysis to follow, we work with $\widetilde{Z}^* U$ and $\widehat{Z} V$ instead of $\widetilde{Z}^*$ and $\widehat{Z}$. To avoid extra notation, from now on, we will use $\widetilde{Z}^*$ and $\widehat{Z}$ for these new versions, which we refer to as *properly aligned*. With this choice, we may assume $U = V = I_r$ in the CS decomposition (45).

The following lemma isolates some useful properties of properly aligned subspaces:

LEMMA 6. *Let $\widetilde{Z}^*$ and $\widehat{Z}$ be properly aligned, and define the matrices*

$$\widehat{P} := P_{\widehat{Z}} - P_{\widetilde{Z}^*} = \widehat{Z}\widehat{Z}^T - \widetilde{Z}^* (\widetilde{Z}^*)^T \quad \text{and} \quad \widehat{E} := \widehat{Z} - \widetilde{Z}^*. \tag{46}$$

In terms of the CS decomposition (45), *we have*:

$$|||\widehat{E}|||_{\mathrm{HS}} \le |||\widehat{P}|||_{\mathrm{HS}}, \tag{47a}$$

$$(\widetilde{Z}^*)^T (P_{\widetilde{Z}^*} - P_{\widehat{Z}}) \widetilde{Z}^* = \widehat{S}^2 \quad \textit{and} \tag{47b}$$

$$\begin{aligned} d^2_{\mathrm{HS}}(\widehat{Z}, \widetilde{Z}^*) &= |||P_{\widetilde{Z}^*} - P_{\widehat{Z}}|||^2_{\mathrm{HS}} \\ &= 2|||\widehat{S}^2|||^2_{\mathrm{HS}} + 2|||\widehat{C}\widehat{S}|||^2_{\mathrm{HS}} \\ &= 2\sum_k \widehat{s}_k^2 (\widehat{s}_k^2 + \widehat{c}_k^2) = 2\,\mathrm{tr}(\widehat{S}^2). \end{aligned} \tag{47c}$$

PROOF. From the CS decomposition (45), we have

$$\widetilde{Z}^*(\widetilde{Z}^*)^T - \widehat{Z}(\widehat{Z})^T = Q \begin{pmatrix} \widehat{S}^2 & -\widehat{C}\widehat{S} & 0 \\ -\widehat{S}\widehat{C} & -\widehat{S}^2 & 0 \\ 0 & 0 & 0 \end{pmatrix} Q^T,$$

from which relations (47b) and (47c) follow. From decomposition (45) and the proper alignment condition $U = V = I_r$, we have

$$\begin{aligned} |||\widehat{E}|||_{\mathrm{HS}}^2 &= |||Q^T(\widehat{Z} - \widetilde{Z}^*)|||_{\mathrm{HS}}^2 = |||I_r - \widehat{C}|||_{\mathrm{HS}}^2 + |||\widehat{S}|||_{\mathrm{HS}}^2 \\ &= 2\sum_{i=1}^{r}(1-\widehat{c}_i) \le 2\sum_{i=1}^{r}(1-\widehat{c}_i^2) = 2\sum_{i=1}^{r}\widehat{s}_i^2 = |||\widehat{P}|||_{\mathrm{HS}}^2, \end{aligned} \tag{48}$$

where we have used the relations $\widehat{C}^2 + \widehat{S}^2 = I_r$, $\widehat{c}_i \in [0,1]$ and $2\operatorname{tr}(\widehat{S}^2) = |||P_{\widetilde{Z}^*} - P_{\widehat{Z}}|||_{\mathrm{HS}}^2$. □

5.2. *Proof of Theorem 1.* Using the notation introduced in Lemma 6, our goal is to bound the Hilbert–Schmidt norm $|||\widehat{P}|||_{\mathrm{HS}}$. Without loss of generality we will assume $s_1 = 1$ throughout. Recalling definition (44) of the random matrix Δ, the following inequality plays a central role in the proof:

LEMMA 7. *Under condition* (A1) *and* $s_1 = 1$, *we have*

$$|||\widehat{P}|||_{\mathrm{HS}}^2 \le 128\langle\langle \widehat{P}, \Delta_1 + \Delta_2 \rangle\rangle \tag{49}$$

with probability at least $1 - \exp(-n/32)$.

PROOF. We use the shorthand notation $\Delta = \Delta_1 + \Delta_2$ for the proof. Since $\widetilde{Z}^*$ is feasible and $\widehat{Z}$ is optimal for problem (20), we have the basic inequality $\langle\langle \widehat{\Sigma}_n, P_{\widetilde{Z}^*} \rangle\rangle \le \langle\langle \widehat{\Sigma}_n, P_{\widehat{Z}} \rangle\rangle$. Using the decomposition $\widehat{\Sigma} = \Gamma + \Delta$ and rearranging yields the inequality

$$\langle\langle \Gamma, P_{\widetilde{Z}^*} - P_{\widehat{Z}} \rangle\rangle \le \langle\langle \Delta, P_{\widehat{Z}} - P_{\widetilde{Z}^*} \rangle\rangle. \tag{50}$$

From definition (44) of Γ and $Z^* = \widetilde{Z}^* R$, the left-hand side of the inequality (50) can be lower bounded as

$$\begin{aligned} \langle\langle \Gamma, P_{\widetilde{Z}^*} - P_{\widehat{Z}} \rangle\rangle &= \langle\langle \overline{B}, SR^T(\widetilde{Z}^*)^T(P_{\widetilde{Z}^*} - P_{\widehat{Z}})\widetilde{Z}^* RS \rangle\rangle \\ &= \operatorname{tr} \overline{B} SR^T \widehat{S}^2 RS \\ &\ge \lambda_{\min}(\overline{B})\lambda_{\min}(S^2)\lambda_{\min}(R^T R)\operatorname{tr}(\widehat{S}^2), \end{aligned}$$

where we have used (90) and (91) of Appendix I several times (cf. the supplementary material [3]). We note that $\lambda_{\min}(S^2) = s_r^2 \ge \frac{1}{2}$ and $\lambda_{\min}(R^T R) \ge \frac{1}{2}$ provided $rC_m(f^*) \ge \frac{1}{2}$; see equation (70). To bound the minimum eigenvalue of $\overline{B}$,

let $\gamma_{\min}(B)$ denote the minimum singular value of the $n \times r$ Gaussian matrix B. The following concentration inequality is well known (cf. [12, 20]):

$$\mathbb{P}\big[\gamma_{\min}(B) \leq \sqrt{n} - \sqrt{r} - t\big] \leq \exp(-t^2/2) \qquad \text{for all } t > 0.$$

Since $\lambda_{\min}(\overline{B}) = \gamma^2_{\min}(B/\sqrt{n})$, we have that $\lambda_{\min}(\overline{B}) \geq (1 - \sqrt{r/n} - t)^2$ with probability at least $1 - \exp(-nt^2/2)$. Assuming $r/n \leq \frac{1}{4}$ and setting $t = \frac{1}{4}$, we get $\lambda_{\min}(\overline{B}) \geq \frac{1}{16}$ with probability at least $1 - \exp(-n/32)$. Putting the pieces together yields the claim. $\square$

Inequality (49) reduces the problem of bounding $|||\widehat{P}|||^2_{\mathrm{HS}}$ to the sub-problem of studying the random variable $\langle\langle \widehat{P}, \Delta_1 + \Delta_2 \rangle\rangle$. Based on Lemma 7, our next step is to establish an inequality (holding with high probability) of the form

$$\langle\langle \widehat{P}, \Delta_1 + \Delta_2 \rangle\rangle \leq c_1 \left\{ \frac{\sigma_m}{\sqrt{n}} r^{3/2} \mathcal{F}\big(|||\widehat{E}|||_{\mathrm{HS}}\big) + \kappa |||\widehat{E}|||^2_{\mathrm{HS}} + \epsilon^2_{m,n} \right\}, \tag{51}$$

where c_1 is some universal constant, κ is the constant in condition (A1) and $\epsilon_{m,n}$ is the critical radius from Theorem 1. Doing so is a nontrivial task: both matrices $\widehat{P}$ and Δ are random and depend on one another, since the subspace $\widehat{Z}$ was obtained by optimizing a random function depending on Δ. Consequently, our proof of bound (51) involves deriving a uniform law of large numbers for a certain matrix class.

Suppose that bound (51) holds, and that the subspaces $\widetilde{Z}^*$ and $\widehat{Z}$ are properly aligned. Lemma 6 implies that $|||\widehat{E}|||_{\mathrm{HS}} \leq |||\widehat{P}|||_{\mathrm{HS}}$, and since $\mathcal{F}$ is a nondecreasing function, inequality (51) combined with Lemma 7 implies that

$$(1 - 128\kappa c_1)|||\widehat{P}|||^2_{\mathrm{HS}} \leq c_1 \left\{ \frac{\sigma_m}{\sqrt{n}} r^{3/2} \mathcal{F}\big(|||\widehat{P}|||_{\mathrm{HS}}\big) + \epsilon^2_{m,n} \right\},$$

from which the claim follows as long as κ is suitably small (e.g., $\kappa \leq \frac{1}{256 c_1}$ suffices). Accordingly, in order to complete the proof of Theorem 1, it remains to prove bound (51), and the remainder of our work is devoted to this goal. Given the linearity of trace, we can bound the terms $\langle\langle \widehat{P}, \Delta_1 \rangle\rangle$ and $\langle\langle \widehat{P}, \Delta_2 \rangle\rangle$ separately.

5.2.1. *Bounding* $\langle\langle \widehat{P}, \Delta_1 \rangle\rangle$. Let $\{\overline{z}_j\}$, $\{\widetilde{z}^*_j\}$ and $\{\widehat{e}_j\}$ and $\{\overline{w}_j\}$ denote the columns of $\widehat{Z}$, $\widetilde{Z}^*$, $\widehat{E}$ and $\overline{W}$, respectively, where we recall the definitions of these quantities from equation (44) and Lemma 6. Note that $\overline{w}_j = n^{-1} \sum_{i=1}^n w_i \beta_{ij}$. In Appendix C.1 of the supplementary material [3], we show that

$$\langle\langle \widehat{P}, \Delta_1 \rangle\rangle \leq \sqrt{6}\sigma r^{3/2} \max_{j,k} |\langle \overline{w}_k, \widehat{e}_j \rangle| + \sqrt{\frac{3}{2}} \sigma r |||\widehat{E}|||^2_{\mathrm{HS}} \max_{j,k} |\langle \overline{w}_j, \widetilde{z}^*_k \rangle|. \tag{52}$$

Consequently, we need to obtain bounds on quantities of the form $|\langle \overline{w}_j, v \rangle|$, where the vector v is either fixed (e.g., $v = \widetilde{z}^*_j$) or random (e.g., $v = \widehat{e}_j$). The following lemmas provide us with the requisite bounds:

LEMMA 8. *We have*

$$\max_{j,k} \sigma r^{3/2} |\langle \overline{w}_k, \widehat{e}_j \rangle| \le C \left\{ \frac{\sigma}{\sqrt{n}} r^{3/2} \mathcal{F}(|||\widehat{E}|||_{\mathrm{HS}}) + \kappa |||\widehat{E}|||_{\mathrm{HS}}^2 + \kappa \epsilon_{m,n}^2 \right\}$$

with probability at least $1 - c_1 r \exp(-\kappa^2 r^{-3} n \frac{\epsilon_{m,n}^2}{2\sigma^2}) - r \exp(-n/64)$.

LEMMA 9. *We have*

$$\mathbb{P}\Big[\max_{j,k} \sigma r |\overline{w}_k^T \widetilde{z}_j^*| \le \sqrt{6}\kappa \Big] \ge 1 - r^2 \exp(-\kappa^2 r^{-2} n / 2\sigma^2).$$

See Appendices C.2 and C.3 in the supplementary material [3] for the proofs of these claims.

5.2.2. *Bounding* $\langle\langle \widehat{P}, \Delta_2 \rangle\rangle$. Recalling definition (44) of Δ_2 and using linearity of the trace, we obtain

$$\langle\langle \widehat{P}, \Delta_2 \rangle\rangle = \frac{\sigma^2}{n} \sum_{j=1}^{r} \{ (\overline{z}_j)^T W^T W \overline{z}_j - (\widetilde{z}_j^*)^T W^T W \widetilde{z}_j^* \}.$$

Since $\widehat{e}_j = \overline{z}_j - \widetilde{z}_j^*$, we have

$$\begin{aligned} \langle\langle \widehat{P}, \Delta_2 \rangle\rangle &= \sigma^2 \sum_{j=1}^{r} \left\{ 2(\widetilde{z}_j^*)^T \left(\frac{1}{n} W^T W - I_r \right) \widehat{e}_j + \frac{1}{n} \|W \widehat{e}_j\|_2^2 + 2(\widetilde{z}_j^*)^T \widehat{e}_j \right\} \\ &\le \sigma^2 \sum_{j=1}^{r} \Big\{ 2 \underbrace{(\widetilde{z}_j^*)^T \left(\frac{1}{n} W^T W - I_r \right) \widehat{e}_j}_{T_1(\widehat{e}_j; \widetilde{z}_j^*)} + \underbrace{\frac{1}{n} \|W \widehat{e}_j\|_2^2}_{T_2(\widehat{e}_j)} \Big\}, \end{aligned} \tag{53}$$

where we have used the fact that $2\sum_j (\widetilde{z}_j^*)^T \widehat{e}_j = 2\sum_j [(\widetilde{z}_j^*)^T \overline{z}_j - 1] = 2\sum_j (\widehat{c}_j - 1) = -|||\widehat{E}|||_{\mathrm{HS}}^2 \le 0$.

The following lemmas provide high probability bounds on the terms T_1 and T_2.

LEMMA 10. *We have the upper bound*

$$\sigma^2 \sum_{j=1}^{r} T_1(\widehat{e}_j; \widetilde{z}_j^*) \le C \left\{ \sigma_0 \frac{\sigma}{\sqrt{n}} r \mathcal{F}(|||\widehat{E}|||_{\mathrm{HS}}) + \kappa |||\widehat{E}|||_{\mathrm{HS}}^2 + \kappa \epsilon_{m,n}^2 \right\}$$

with probability $1 - c_2 \exp(-\kappa^2 r^{-2} n \frac{\epsilon_{m,n} \wedge \epsilon_{m,n}^2}{16\sigma^2}) - r \exp(-n/64)$.

LEMMA 11. *We have the upper bound* $\sigma^2 \sum_{j=1}^{r} T_2(\widehat{e}_j) \le C\kappa \{ |||\widehat{E}|||_{\mathrm{HS}}^2 + \epsilon_{m,n}^2 \}$ *with probability at least* $1 - c_3 \exp(-\kappa^2 r^{-2} n \epsilon_{m,n}^2 / 2\sigma^2)$.

See Appendices C.4 and C.5 in the supplementary material [3] for the proofs of these claims.

6. Discussion. We studied the problem of sampling for functional PCA from a functional-theoretic viewpoint. The principal components were assumed to lie in some Hilbert subspace $\mathcal{H}$ of L^2, usually a RKHS, and the sampling operator, a bounded linear map $\Phi : \mathcal{H} \to \mathbb{R}^m$. The observation model was taken to be the output of Φ plus some Gaussian noise. The two main examples of Φ considered were time sampling, $[\Phi f]_j = f(t_j)$ and (generalized) frequency truncation $[\Phi f]_j = \langle \psi_j, f \rangle_{L^2}$. We showed that it is possible to recover the subspace spanned by the original components, by applying a regularized version of PCA in $\mathbb{R}^m$ followed by simple linear mapping back to function space. The regularization involved the "trace-smoothness condition" (19) based on the matrix $K = \Phi\Phi^*$ whose eigendecay influenced the rate of convergence in $\mathbb{R}^m$.

We obtained the rates of convergence for the subspace estimators both in the discrete domain, $\mathbb{R}^m$, and the function domain, L^2. As examples, for the case of a RKHS $\mathcal{H}$ for which both the kernel integral operator and the kernel matrix K have polynomial-α eigendecay (i.e., $\mu_j \asymp \widehat{\mu}_j \asymp j^{-2\alpha}$), the following rates in HS-projection distance for subspaces in the function domain were worked out in detail:

Time sampling	Frequency truncation
$(\frac{1}{mn})^{2\alpha/(2\alpha+1)} + (\frac{1}{m})^{2\alpha}$	$(\frac{1}{n})^{2\alpha/(2\alpha+1)} + (\frac{1}{m})^{2\alpha}$

The two terms in each rate can be associated, respectively, with the estimation error (due to noise) and approximation error (due to having finite samples of an infinite-dimensional object). Both rates exhibit a trade-off between the number of statistical samples (n) and that of functional samples (m). The two rates are qualitatively different: the two terms in the time sampling case interact to give an overall fast rate of n^{-1} for the optimal trade-off $m \asymp n^{1/(2\alpha)}$, while there is no interaction between the two terms in the frequency truncation; the optimal trade-off gives an overall rate of $n^{-2\alpha/(2\alpha+1)}$, a characteristics of nonparametric problems. Finally, these rates were shown to be minimax optimal.

SUPPLEMENTARY MATERIAL

Proofs and auxiliary results (DOI: 10.1214/12-AOS1033SUPP; .pdf). This supplement contains some of the proofs and auxiliary results referenced in the text.

REFERENCES

[1] AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

[2] AMINI, A. A. and WAINWRIGHT, M. J. (2012). Approximation properties of certain operator-induced norms on Hilbert spaces. *J. Approx. Theory* **164** 320–345. MR2864648

[3] AMINI, A. A. and WAINWRIGHT, M. J. (2012). Supplement to "Sampled forms of functional PCA in reproducing kernel Hilbert spaces." DOI:10.1214/12-AOS1033SUPP.
[4] BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, MA. MR2239907
[5] BESSE, P. and RAMSAY, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika* **51** 285–311. MR0848110
[6] BHATIA, R. (1996). *Matrix Analysis*. Springer, New York.
[7] BOENTE, G. and FRAIMAN, R. (2000). Kernel-based functional principal components. *Statist. Probab. Lett.* **48** 335–345. MR1771495
[8] BOSQ, D. (2000). *Linear Processes in Function Spaces*: *Theory and Applications*. *Lecture Notes in Statistics* **149**. Springer, New York. MR1783138
[9] CAI, T. T. and YUAN, M. (2010). Nonparametric covariance function estimation for functional and longitudinal data. Technical report, Georgia Institute of Technology.
[10] CARDOT, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonparametr. Stat.* **12** 503–538. MR1785396
[11] 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
[12] DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces*, *Vol. I* 317–366. North-Holland, Amsterdam. MR1863696
[13] DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007
[14] GU, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York. MR1876599
[15] HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 109–126. MR2212577
[16] 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
[17] HUANG, J. Z., SHEN, H. and BUJA, A. (2008). Functional principal components analysis via penalized rank one approximation. *Electron. J. Stat.* **2** 678–695. MR2426107
[18] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961
[19] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448
[20] LEDOUX, M. (2001). *The Concentration of Measure Phenomenon*. *Mathematical Surveys and Monographs* **89**. Amer. Math. Soc., Providence, RI. MR1849347
[21] 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
[22] MENDELSON, S. (2002). Geometric parameters of kernel machines. In *Computational Learning Theory* (*Sydney*, 2002). *Lecture Notes in Computer Science* **2375** 29–43. Springer, Berlin. MR2040403
[23] PAUL, D. and JOHNSTONE, I. (2008). Augmented sparse principal component analysis for high-dimensional data. Available at arXiv:1202.1242.
[24] PEZZULLI, S. and SILVERMAN, B. W. (1993). Some properties of smoothed principal components analysis for functional data. *Comput. Statist.* **8** 1–16. MR1220336
[25] QI, X. and ZHAO, H. (2010). Functional principal component analysis for discretely observed functional data. Unpublished manuscript.
[26] RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*: *Methods and Case Studies*. Springer, New York. MR1910407

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

[28] RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **53** 233–243. MR1094283

[29] SILVERMAN, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.* **24** 1–24. MR1389877

[30] VAN DE GEER, S. A. (2009). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

[31] WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442

[32] 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

DEPARTMENT OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: amini@eecs.berkeley.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: wainwrig@stat.berkeley.edu