Electronic Journal of Statistics
Vol. 17 (2023) 156–197
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2096

# Optimal function-on-scalar regression over complex domains*

**Matthew Reimherr and Bharath Sriperumbudur**

*Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA*
*e-mail:* mreimherr@psu.edu; bks18@psu.edu

**Hyun Bin Kang**

*Department of Statistics, Western Michigan University, Kalamazoo, MI 49008, USA*
*e-mail:* h.kang@wmich.edu

**Abstract:** In this work we consider the problem of estimating function-on-scalar regression models when the functions are observed over multi-dimensional or manifold domains and with potentially multivariate output. We establish the minimax rates of convergence and present an estimator based on reproducing kernel Hilbert spaces that achieves the minimax rate. To better interpret the derived rates, we extend well-known links between RKHS and Sobolev spaces to the case where the domain is a compact Riemannian manifold. This is accomplished using an interesting connection to Weyl's Law from partial differential equations. We conclude with a numerical study and an application to 3D facial imaging.

**Keywords and phrases:** Functional data analysis, reproducing kernel Hilbert space, functional regression, optimal regression, Weyl's law.

Received November 2021.

## Contents



*M. Reimherr was supported by NSF grant SES 1853209. B. Sriperumbudur was supported by NSF grant DMS 1945396.

## 1. Introduction

Functional data analysis has seen a precipitous development in recent decades, in terms of methodology, theory, and applications. As with classical statistics, functional linear regression models are used extensively in practice. In recent years, there has also been a surge in the development of so-called *next generation functional data analysis*, which involves functional data with highly complex structures. Much of this development has been spurred by advances in biomedical imaging, where dense measurements are taken over various tissues, including the brain, arteries, eyes, and faces (e.g., Ettinger et al., 2016, Lila et al., 2016, Kang et al., 2017, Choe et al., 2017, Lee et al., 2018). In each of these examples, the measurements are taken over complex spatial domains such as $\mathbb{R}^3$ or two-dimensional manifolds.

Establishing the optimality of parameter estimates in FDA remains an important topic given the complexity of the data and models involved. Indeed, depending on the problem, one can see a wide variety of convergence rates. For example, in univariate mean estimation it was shown that the rates depend on the smoothness of the underlying parameter as well as the sampling frequency of the data; depending on how often the functions are sampled, one can obtain a parametric convergence rate or nonparametric convergence rate (Cai and Yuan, 2011, Li et al., 2010, Zhang et al., 2016). In scalar-on-function regression, the rates relate both to the smoothness of the slope function and the regularity of the predictor function; these rates have been extended to nonlinear models as well (Hall et al., 2007, Cai and Yuan, 2012, Wang and Ruppert, 2015, Reimherr et al., 2017, Sun et al., 2018). In high-dimensional function-on-scalar regression models it was shown that the convergence rates match the classic scalar outcome setting as long as the sampling is dense enough (Barber et al., 2017, Fan and Reimherr, 2017). In principal component estimation, one obtains rates that reflect how deep into the spectrum one wishes to estimate as well as how spread out the eigenvalues are (Dauxois et al., 1982, Jirak, 2016, Petrovich and Reimherr, 2017). In each of these cases, different rates can be obtained depending on the regularity of the problem. However, optimality of function-on-scalar regression, especially with more complex domains and sampling schemes, has not yet been established. This problem is most similar to the mean estimation since the predictors are scalars, though the estimation error depends on the number of predictors. Such results are critical given the recent developments of functional data methods involving manifolds (Kang et al., 2017, Dai et al.,

2018, Lin and Yao, 2018). In this work we address this issue by: (1) establishing minimax lower bounds on the estimation rate (2) providing a minimax optimal estimator whose upper bounds match the developed lower bounds and (3) interpreting the rate via a new connection to Sobolev spaces over manifold domains.

We develop our theory under a fairly general structure:

$$Y_{ij\ell} = Y_{i\ell}(u_{ij}) + \delta_{ij\ell} = \sum_{p=1}^{P} X_{ip}\beta_{\ell p}(u_{ij}) + \varepsilon_{i\ell}(u_{ij}) + \delta_{ij\ell}, \tag{1.1}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, and $\ell = 1, \ldots, L$. Here $i$ indexes the subject, $j$ the observed domain point, and $\ell$ the coordinates of the functional outcomes. Intuitively, this means that for each subject we have $L$ functional outcomes, $Y_{i\ell}(u) \in \mathbb{R}$, that are only observed at domain points $u_{ij} \in \mathcal{U}$, and $P$ scalar predictors, $X_{ip}$. The random component is decomposed into a smooth subject specific error, $\varepsilon_{i\ell}$ and a noise, $\delta_{ij\ell}$, both of which are assumed to be independent across $i$ and $j$, though potentially dependent across $\ell$. The domain $\mathcal{U}$ is most commonly the interval $[0, 1]$, but it may also be a more complex manifold, both of which are included in our theory. For example, in Ettinger et al. (2016) $Y_{ijl}$ represents the thickness of the internal carotid artery, meaning that $\mathcal{U}$ is a two dimensional manifold representing the artery and sits in a three dimensional space with $L = 1$ (since only the thickness is measured); predictors, $X_{ip}$, of interest include age, weight, smoking status, etc. The $\varepsilon_{i\ell}$ represents the individual level variation that is not accounted for by the predictors, while $\delta_{ij\ell}$ represents measurement noise. In Kang et al. (2017) they consider the shape of human faces and $Y_{ij\ell}$ represents the position of the face in 3D space. In their framework $\mathcal{U}$ is taken as a common reference face resulting in $\mathcal{U}$ being a two dimensional manifold while $L = 3$ since the face is measured in three dimensions. The predictors in their application include age, weight, and genetic measurements.

This paper is concerned with optimal estimation rates of the unknown parameter functions, $\beta_{\ell p} : \mathcal{M} \to \mathbb{R}$. The intrinsic dimension of $\mathcal{U}$ plays a critical role in determining the *phase transition* of the minimax estimation rates (that is, the point where one reaches a parametric rate) for $\beta_{\ell k}(u)$, while, interestingly, the values $L$ and $P$ do not. In addition, it was previously thought that, in simpler settings, such as mean estimation, it was necessary to control the smoothness of the underlying functions $Y_{i\ell}(u)$ (Cai and Yuan, 2011), or equivalently the errors $\varepsilon_{ij}(u)$, however, we show that this is actually unnecessary and establish all of our results under the mild assumption that $\sup_{u\in\mathcal{U}} \mathrm{Var}(\varepsilon_{i\ell}(u)) < \infty$, that is, the point-wise variance of the errors is bounded.

We assume that $\beta_{\ell p}$ (for all $\ell, p$) lie in a reproducing kernel Hilbert space (RKHS), and establish our rates relative to the rate of decay of the eigenvalues of the kernel defining the RKHS. In contrast, Cai and Yuan (2011) develop theory for one dimensional mean estimation assuming the parameters lie in a particular Sobolev space, which will be included in our theory as a special example. Under

mild assumptions, we will show that the optimal rate of convergence is given by

$$O_P\left(LP\left[(nm)^{-\frac{2h}{2h+1}}+n^{-1}\right]\right),$$

where $h$ is related to the kernel of the RKHS, $n$ is the sample size, and $m$ is the harmonic mean of the $m_i$. In Section 4.3, we consider the case where $\mathcal{U}$ is a compact $d$-dimensional Riemannian manifold. When the parameters $\beta_{\ell k}$ posses $r$ derivatives, we use a connection with Weyl's law to show that $h = r/d$, which extends well known results for Sobolev spaces on $\mathbb{R}^d$ (Edmunds and Triebel, 1996) resulting in the rate

$$O_P\left(LP\left[(nm)^{-\frac{2r}{2r+d}}+n^{-1}\right]\right),$$

which clearly shows the effect of the intrinsic dimension of $\mathcal{U}$ on the convergence rates of our estimators, with higher dimensions leading to slower rates. This further highlights the utility in exploiting manifold structures that may reside in higher dimensional spaces; the convergence rate is tied only to the intrinsic dimension of the manifold and not to that of the ambient space.

The remainder of the paper is organized as follows. In Section 2 we provide an overview of the modeling assumptions and necessary mathematical tools. In Section 3 we define our estimation procedure and provide a formulation useful for establishing mathematical properties. In Section 4 we collect our theoretical contributions, which constitute the primary novel contributions of the paper. There we provide a general lower bound on the minimax rate, followed by a theorem showing that our proposed estimator achieves the optimal rate. We conclude the section with discussion on the derived rate. We provide a new connection between the eigenvalues of an RKHS and Sobolev spaces over manifold domains, which allow us to interpret our results in terms of the dimension of the domain and the smoothness of the parameters being estimated. We conclude the paper with numerical work in Section 5, where we provide simulations that further articulate the rates seen in Section 4. We also provide an application to 3D facial imaging from anthropology, highlighting the utility of such tools in biomedical imaging.

## 2. Background and modeling assumptions

Here we provide the necessary background as well as a clear outline of our modeling assumptions.

### *2.1. Reproducing kernel Hilbert spaces*

RKHSs provide a variety of benefits for functional data analysis. The first is that the kernel can be tailored to reflect certain beliefs or assumptions about the parameters, e.g., smoothness or periodicity. The second is that the eigenfunctions of the kernel can be used as a basis for approximating functional observations

and/or parameter estimates, though the reproducing property can also be used to obtain parameter estimates. Lastly, commonly used spaces, such as Sobolev spaces, as well as estimation techniques such as smoothing splines can naturally be viewed in an RKHS framework (Wahba, 1990, Berlinet and Thomas-Agnan, 2011).

We assume throughout that $\mathcal{U}$ is a compact $d$-dimensional manifold with $d < \infty$, i.e., $\mathcal{U}$ is a second countable compact Hausdorff space such that each point $u \in \mathcal{U}$ is contained in an open set that is homeomorphic to an open set in $\mathbb{R}^d$. We assume that $\mathcal{U}$ is equipped with a countably additive measure, $\mu$, with respect to the Borel $\sigma$-algebra, whose support is $\mathcal{U}$ and satisfies $\mu(\mathcal{U}) = 1$. This means that we can define integrals over $\mathcal{U}$ and the space, $L^2(\mathcal{U}, \mu)$, of square integrable functions over $\mathcal{U}$ is equipped with the inner product

$$\langle f, g \rangle = \int_{\mathcal{U}} f(u)g(u)\, d\mu(u).$$

Throughout, for notational simplicity, we will often write $L^2$ for $L^2(\mathcal{U}, \mu)$. A kernel function, $K : \mathcal{U} \times \mathcal{U} \to \mathbb{R}^+$, is a bivariate function that is symmetric, positive definite, and continuous (though this can be relaxed). There is a one-to-one correspondence between RKHSs and kernel functions. One can generate the RKHS from $K$ in at least one of two ways, though for our purposes one in particular is especially useful (Berlinet and Thomas-Agnan, 2011, Section 3.2). Note that any norm $\|\cdot\|$ or inner product $\langle \cdot, \cdot \rangle$ written without subscript is understood to be with respect to $L^2$. By Mercer's theorem we can write

$$K(u, s) = \sum_{k=1}^{\infty} \tau_k v_k(u) v_k(s),$$

where $v_k \in L^2$ are orthonormal and $\{\tau_k\}$ is a positive, non-increasing, summable sequence, with the convergence holding in an absolute and uniform sense. One can then obtain $\mathbb{K}$ as the subset

$$\mathbb{K} = \left\{ f \in L^2 : \sum_{k=1}^{\infty} \frac{\langle f, v_k \rangle^2}{\tau_k} < \infty \right\}.$$

Then $\mathbb{K}$ is an RKHS when equipped with the inner product $\langle f, g \rangle_{\mathbb{K}} = \sum_k \tau_k^{-1} \langle f, v_k \rangle \langle g, v_k \rangle$. On a technical note, since $L^2$ is a set of equivalence classes one is implicitly taking $f \in \mathbb{K}$ to be the unique member of each class that is continuous. This view is especially useful as it emphasizes how quickly the coordinates of $f$ must decay when expressed in the $\{v_j\}$ basis, which is critical for understanding and developing minimax rates.

In general, constructing RKHS kernels (which must be positive definite) over manifolds is a nontrivial task. In particular, simply replacing a Euclidean distance within a kernel with a Riemannian distance will not, in general, result in a valid RKHS kernel (Jayasumana et al., 2013). In our applications we consider manifolds that can be covered by a single chart, which makes the problem simpler as we can map the problem to Euclidean space to build valid kernels, as illustrated in Section 5.

### *2.2. Modeling assumptions*

We now state our modeling assumptions. We provide a summary at the end of this section for ease of reference. The model in (1.1) assumes that the underlying trajectories are assumed at a small number of points and with error. The parameters, $\beta_{\ell k}$ are assumed to lie within $\mathbb{K}$. Regularity assumptions about $\beta_{\ell k}$ are introduced by making assumptions about $\mathbb{K}$, especially the rate at which the eigenvalues of $K$ converge to zero.

Unlike in Cai and Yuan (2011), we make only minimal assumptions about the regularity of $\varepsilon_{i\ell}(u)$. In particular, we establish our minimax rates under the mild assumption that the point-wise variance of the errors is bounded, $\sup_{u\in\mathcal{U}} \operatorname{Var}(\varepsilon_{i\ell}(u)) < \infty$, which implies (and is only slightly stronger than) $\mathrm{E}\,\|\varepsilon_{i\ell}\|^2 < \infty$. In Cai and Yuan (2011) the much stronger assumption that the errors are in the RKHS, $\mathrm{E}\,\|\varepsilon_{i\ell}\|_{\mathbb{K}}^2 < \infty$, was made, which, by the reproducing property implies our assumption. While seemingly innocent, this is an incredibly strong assumption that would actually preclude achieving optimal convergence rates in most settings. Practically, the data is usually much rougher than the underlying mean parameters. However, requiring that they reside in the same space implies that the $\beta_{\ell p}$ can only be smoothed up to the smoothness of the data. For example, if $\mathcal{U} = [0,1]$ and $\beta_{\ell p}$ possessed two derivatives, while $\varepsilon_{i\ell}$ only possessed one, then the rate given by Cai and Yuan (2011) would be $(nm)^{-2/3} + n^{-1}$, however, as we will show, this rate can be improved to $(nm)^{-4/5} + n^{-1}$. Furthermore, in settings such as finance or geosciences, $\varepsilon_{i\ell}$ might not possess any derivatives or be part of any RKHS (e.g. Brownian motion or the Ornstein-Uhlenbeck process).

We treat the predictors, $X_{ij}$, as deterministic. The observed points $u_{ij}$ will be assumed to be iid draws from $\mathcal{U}$, with density (w.r.t. $\mu$) that is bounded away from 0 and $\infty$. We also assume that the functional outcome is observed with error, namely $Y_{ij\ell} = Y_{i\ell}(u_{ij}) + \delta_{ij\ell}$. The errors $\delta_{ij\ell}$ are assumed to be iid across $i$ and $j$, though they can be dependent in $\ell$. We assume these errors are centered and have finite variance. We now summarize all of the assumptions introduced in this section.

**Assumption 2.1.** *We make the following modeling assumptions.*

1. *The observed data are* $\{Y_{ij\ell}, u_{ij}, X_{i1}, \ldots, X_{iP}\}$ *for* $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, *and* $\ell = 1, \ldots, L$, *with* $Y_{ij\ell} \in \mathbb{R}$ *and* $X_{ip} \in \mathbb{R}$.
2. *The observed domain locations,* $u_{ij}$, *are iid elements of* $\mathcal{U}$, *a compact* $d$*-dimensionl manifold. The space* $\mathcal{U}$ *is equipped with a countably additive measure* $\mu$ *(over the Borel* $\sigma$*-field) with* $\mu(\mathcal{U}) = 1$. *The random elements* $u_{ij}$ *are assumed to have a density (w.r.t.* $\mu$*) which is bounded above and below (from 0).*
3. *The observed data satisfy the linear model*

$$Y_{ij\ell} = \sum_{p=1}^{P} X_{ip}\beta_{\ell p}(u_{ij}) + \varepsilon_{i\ell}(u_{ij}) + \delta_{ij\ell}.$$

4. *The mean parameters reside within the RKHS, i.e.,* $\beta_{\ell p} \in \mathbb{K}$*, with continuous kernel* $K(u,s)$*. The eigenvalues of* $K$ *satisfy* $\tau_k \asymp k^{-2h}$ *for* $h \geq 1$*.*
5. *The sequences* $\varepsilon_{i\ell} \in L^2$*,* $u_{ij} \in \mathcal{U}$*, and* $\delta_{ij\ell}$ *are random and independent of each other.*
6. *The covariates* $X_{ip}$ *are deterministic. Define* $\Sigma_X = n^{-1}\sum \mathbf{X}_i\mathbf{X}_i^\top$*, where* $\mathbf{X}_i = (X_{i1},\ldots,X_{iP})$*, and assume that smallest and largest eigenvalues are bounded away from* 0 *for all large* $n$*:* $0 < \nu^{-1} \leq \sigma_{\min}(\Sigma_X) \leq \sigma_{\max}(\Sigma_X) \leq \nu < \infty$*.*
7. *Assume that the norm of the predictors are bounded* $|\mathbf{X}_i|^2 \leq \zeta < \infty$*.*
8. *The* $\delta_{ij\ell}$ *represent the measurement error and are iid across* $i$ *and* $j$*, though potentially dependent across* $\ell$*. They have mean zero and finite variance,* $\mathrm{Var}(\delta_{ij\ell}) \leq M_\delta < \infty$*, for some fixed* $M_\delta \in \mathbb{R}$*.*
9. *The stochastic processes* $\varepsilon_{i\ell}$ *are iid across* $i$*, though potentially dependent across* $\ell$*. They are assumed to have mean zero and to satisfy* $\sup_{u\in\mathcal{U}} \mathrm{Var}(\varepsilon_{i\ell}(u)) \leq M_\epsilon < \infty$*, for some fixed* $M_\epsilon \in \mathbb{R}$*.*

## 3. Estimation methodology

We assemble an estimate of each $\ell$ coordinate separately. Define the $\ell$th target function as

$$\begin{aligned}\mathcal{L}^\ell_{mn}(\mathbf{b}) &= \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i}(Y_{ijl} - \mathbf{X}_i^\top \mathbf{b}(u_{ij}))^2 + \lambda\sum_{k=1}^p \|b_k\|^2_{\mathbb{K}} \\ &= \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i}(Y_{ijl} - \langle \mathbf{X}_i^\top \mathbf{b}, K_{u_{ij}}\rangle_{\mathbb{K}})^2 + \lambda\|\mathbf{b}\|^2_{\mathbb{K}},\end{aligned}$$

where $b_k \in \mathbb{K}$ and $\mathbf{b} = (b_1,\ldots,b_P)$ are the generic arguments of the target function and $\mathbf{X}_i = (X_{i1},\ldots,X_{iP})$ are the covariates for the *ith* unit. While increased performance can be gained by considering the $L$ coordinates jointly, it will not impact the minimax rates. The minimizer $\hat{\boldsymbol{\beta}}_\ell$, can be obtained in a closed form using operator notation (as opposed to the representer theorem). We can take the derivative with respect to $\mathbf{b}$ (in the $\mathbb{K}$ topology) as

$$D\mathcal{L}^\ell_{mn}(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i} -2(Y_{ij\ell} - \langle K_{u_{ij}}, \mathbf{X}_i^\top \mathbf{b}\rangle_{\mathbb{K}})K_{u_{ij}}\mathbf{X}_i + 2\lambda\mathbf{b},$$

where $K_{u_{ij}}(u) := K(u_{ij},u)$. Define $\mathbf{h}_{nm\ell} \in \mathbb{K}^p$ as

$$\mathbf{h}_{nm\ell} = \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i} Y_{ij\ell}K_{u_{ij}}\mathbf{X}_i, \tag{3.1}$$

and the linear operator $\mathbf{T}_{nm} : \mathbb{K}^p \to \mathbb{K}^p$ as

$$\mathbf{T}_{nm}(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i} \mathbf{X}_i\mathbf{X}_i^\top \mathbf{f}(u_{ij})K_{u_{ij}} = \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i}\langle \mathbf{X}_i^\top \mathbf{f}, K_{u_{ij}}\rangle_{\mathbb{K}}K_{u_{ij}}\mathbf{X}_i. \tag{3.2}$$

Setting the derivative equal to zero we get the operator form for the estimator

$$D\mathcal{L}^{\ell}_{mn}(\mathbf{b}) = -2\mathbf{h}_{nm\ell} + 2\mathbf{T}_{mn}\mathbf{b} + 2\lambda\mathbf{b} = 0 \Longrightarrow \hat{\boldsymbol{\beta}}_{\ell} = (\mathbf{T}_{nm} + \lambda\mathbf{I})^{-1}\mathbf{h}_{nm\ell}.$$

This operator form for $\hat{\boldsymbol{\beta}}_{\ell}$ is convenient for asymptotic theory. In Section 5.1 we will discuss an efficient computational approximation to $\hat{\boldsymbol{\beta}}_{l}$. Using the representer theorem for RKHS, it is also possible to obtain an alternative equivalent formulation for $\hat{\boldsymbol{\beta}}_{l}$ that can be computed exactly, but requires solving large systems of linear equations that can make it impractical for larger datasets.

## 4. Theoretical results

We now provide our key theoretical results. The first is a lower bound on the best possible estimation rate. This bound is obtained using an application of Fano's lemma. Second, we provide an estimator whose upper bound matches the lower bound, implying that it is optimal in a minimax sense. Lastly, we provide an interpretation of the resulting rate by making a connection to Sobolev spaces with domains consisting of compact Riemannian manifolds.

### *4.1. Lower bound*

Recall that when referring to a minimax rate, we have to specify the loss function as well as the class of models we are considering. Here, our loss is based on the $L^2(\mathcal{U}, \mu)$ norm, and we consider all models as outlined in Assumption 2.1. A more delicate point is that we should also specify which quantities in the problem are "fixed", that is, which quantities should be treated as fixed when constructing the scenario that achieves the desired lower bound. This is important since our problem is regression and we are treating the predictors as fixed. So consider $\mathfrak{M}$ to be the collection of all possible distributions for $\{Y_{ij\ell}\}$ for a fixed set of predictors $\{X_{ik}\}$ and fixed $m_i$ (though the $m_i$ are still allowed to vary with $n$) satisfying Assumption 2.1. We also assume that the parameters of interest lie in a closed bounded ball of $\mathbb{K}$: $\|\beta_{\ell p}\|_{\mathbb{K}} \leq M_0$, which will be denoted as $B_{\mathbb{K}}$. So each $\mathcal{M} \in \mathfrak{M}$ indicates the distributions for $(\epsilon_{i\ell}, \delta_{ij\ell}, u_{ij})$ and specifies the values of $\beta_{\ell p}$. Define the estimation error:

$$R_n = \sum_{p=1}^{P}\sum_{\ell=1}^{L} \|\hat{\beta}_{\ell p} - \beta_{\ell p}\|^2.$$

We say that the rate of convergence of $\hat{\boldsymbol{\beta}}$ is $a_n$ if $R_n = O_P(a_n)$. The minimax estimation risk is then defined as the optimal rate of convergence (i.e., the smallest $a_n$), across all possible estimators, in the worst case modeling scenario.

**Theorem 4.1.** *Let $\mathfrak{M}$, as described above, be the collection of probability models satisfying Assumption 2.1 with $\|\beta_{\ell p}\|_{\mathbb{K}} \leq M_0 < \infty$ for all $\ell, p$. Then for any $\hat{\boldsymbol{\beta}}$ which is a function of the data, the estimation error satisfies*

$$\limsup_{n\to\infty} \sup_{\mathcal{M}\in\mathfrak{M}} P(R_n \leq \epsilon LP((nm)^{-2h/(2h+1)} + n^{-1})) \to 0 \qquad \text{as } \epsilon \to 0,$$

*if the arithmetic and harmonic means of the $m_i$ are asymptotically equivalent and the eigenvalues, $\tau_k$, of $K$, decay as $\tau_k \asymp k^{-2h}$.*

The proof of Theorem 4.1 is given in the appendix. It shows that no estimator can achieve a "worst case" rate faster than $LP[(nm)^{-2h/(2h+1)} + n^{-1}]$; we will show in the next section that this bound is tight by giving an estimator that achieves the lower bound. The proof is based on an application of Fano's lemma. We show that a sequence of parameters within the ball $B_{\mathbb{K}}$ can be selected which are sufficiently far apart with respect to the $\mathbb{K}$-norm. We then prove a bound on the Kullback-Leibler divergence between any pair of probability measures induced by this collection of parameters. Combining these two bounds, we are able to apply Fano's lemma to obtain the desired result.

One interesting caveat to Theorem 4.1 is the requirement that the arithmetic and harmonic means of the $m_i$ be asymptotically equivalent. This is due to the arguments of the upper bound being in terms of the harmonic mean and the lower bound arguments in term of the arithmetic mean; this is not simply a theoretical convenience as a case where this doesn't hold becomes surprisingly delicate. For example, suppose that one of the $m_i$ was essentially infinite (implying the entire curve is observed). Then the arithmetic mean would be infinite, but the convergence rate need not be parametric. Alternatively, if even of a small fraction of the $m_i$ were infinite (or very large), then the rate would become parametric, however the harmonic mean need not be infinite especially if the remaining $m_i$ are small. If one were to let the fraction of $m_i$ being infinite (or very large) change with $n$, then one could obtain basically any convergence rate desired (between nonparametric and parametric), all while maintaining a bounded harmonic mean and an infinite arithmetic mean. To avoid this, the lower bound given in Cai and Yuan (2011) was also taken over all $m_i$ that satisfy a specific harmonic mean, however this is somewhat strange since the $m_i$ are actually observed in a given problem. Recently Zhang and Wang (2018) discussed optimal weighting as a function of the $m_i$ in the context of mean and covariance function estimation. However, the weights were chosen to optimize the asymptotic upper bound of a local linear smoother and depended on the choice of the smoothing parameter.

### *4.2. Upper bound*

Recall that our proposed estimator is given by

$$\hat{\boldsymbol{\beta}}_\ell = (\mathbf{T}_{nm} + \lambda \mathbf{I})^{-1} \mathbf{h}_{nm\ell}. \tag{4.1}$$

We first give a more general result that provides a deeper understanding of the components of the convergence rate.

**Theorem 4.2.** *Assume that Assumption 2.1 holds and that $\hat{\boldsymbol{\beta}}$ is as given in (4.1). If $\lambda$ is such that $nm\lambda^{\delta+1/2h} \to \infty$ for some $\delta > 1/2h$, then the estimation error satisfies*

$$R_n = O_P(1) LP\left[\lambda + \frac{1}{\lambda^{1/2h} nm} + \frac{1}{n}\right].$$

Here we see three core components driving the statistical properties of $\hat{\boldsymbol{\beta}}$. As is common in nonparametric smoothing, the bias is given by $\lambda$. The stochastic error is driven by two components. The first is driven by the total number of observed values and takes a familiar "nonparametric rate." The last component is a parametric rate, but only decreases with $n$, reflecting that there is a bounded amount of information that can be extracted from a single function/unit. Balancing the bias and stochastic error, we arrive at the optimal rate of convergence.

**Theorem 4.3.** *Assume that Assumption 2.1 holds and that* $\hat{\boldsymbol{\beta}}$ *is as given in* (4.1). *If* $\lambda \asymp (nm)^{-2h/(1+2h)}$ *then the estimation error satisfies*

$$\limsup_{n\to\infty} \sup_{\beta_{\ell k}\in B_{\mathbb{K}}} P(R_n \geq \epsilon^{-1} L P((nm)^{-2h/(2h+1)} + n^{-1})) \to 0 \qquad \text{as } \epsilon \to 0.$$

Combining Theorems 4.1 and 4.3 we get that the minimax rate of convergence is $(nm)^{-2h/(2h+1)} + n^{-1}$. Furthermore, this rate holds quite broadly across different $\mathbb{K}$. The *phase-transition* occurs when the rate becomes parametric, i.e., $n^{-1}$. Clearly this occurs if

$$(nm)^{-2h/(2h+1)} \ll n^{-1} \iff m \gg n^{1/2h}.$$

In other words, the rate becomes parametric if the (harmonic) average number of points per curve is more than $n^{1/2h}$. If $m$ is less, then the rate is slower than parametric. In the worst case, when $m$ is bounded, the rate becomes the classic nonparametric rate $n^{-2h/(2h+1)}$.

### *4.3. Interpreting the rate*

In our theory, $h$ is only tied to the rate of decay of the eigenvalues of the reproducing kernel. However, there are settings where this rate can be made more interpretable. In the remainder of this section, we state the following theorem for Riemannian manifolds, which ties together several classic results from nonlinear analysis, and extends well-known connections between RKHS and Sobolev spaces for Euclidean spaces. As the proof uses a number of results that might be of interest to readers, we state it here instead of in the appendix.

**Theorem 4.4.** *Let* $\mathcal{U}$ *be a compact* $d$*-dimensional Riemannian manifold. Let* $H^r(\mathcal{U})$ *denote the Sobolev space of real valued functions whose first* $r$ *weak derivatives are in* $L^2(\mathcal{U})$ *and assume* $2r > d$. *Then* $H^r(\mathcal{U})$ *is a reproducing kernel Hilbert space and the eigenvalues of the reproducing kernel decay like* $\tau_k \asymp k^{-2r/d}$.

*Proof.* The Sobolev space, $\mathcal{H}^r := H^r(\mathcal{U})$, of real functions over $\mathcal{U}$ with $r$ weak derivatives in $L^2(\mathcal{U})$ can be continuously embedded into the space of continuous functions, $C(\mathcal{U})$, if $2r > d$ (Hebey, 2000, Section 2.3). This means that we can identify each $f \in \mathcal{H}^r$ as the unique continuous representative of its corresponding equivalence class. This also implies that $\|f\|_{C(\mathcal{U})} \leq M\|f\|_{\mathcal{H}^r}$, for some $M > 0$

(across all $f$). Since $f(x) \leq \|f\|_{C(\mathcal{U})}$ this means point-wise evaluation would be a continuous linear functional on $\mathcal{H}^r$ and by the Riesz representation theorem, the space must also be an RKHS (recall a Hilbert space where point-wise evaluations are continuous is necessarily an RKHS).

One can construct a kernel function that gives rise to $\mathcal{H}^r$ using the Laplace-Beltrami operator (i.e. the Laplacian for manifolds) acting over the space of infinitely differentiable functions, $\Delta : C^\infty(\mathcal{U}) \to C^\infty(\mathcal{U})$. This operator has eigenvalues tending to infinity, which we will label $0 \leq \xi_1 \leq \xi_2 \leq \dots$ (and can be zero), and corresponding eigenfunctions $v_1, v_2, \dots$, which, while infinitely differentiable, can be taken to be an orthonormal basis of $L^2(\mathcal{U})$ (Canzani, 2013, Section 7.1). The Sobolev space $\mathcal{H}^r$ can be identified as

$$\mathcal{H}^r = \left\{ f \in L^2(\mathcal{U}) : \sum_{k=1}^{\infty} \xi_k^r \langle v_k, f \rangle^2 < \infty \right\},$$

see, e.g., Chapter 3 of Craioveanu et al. (2013). Note that the first eigenvalue, $\xi_1$ is usually zero, meaning we do not restrict a function $f$ in that direction. We can equip $\mathcal{H}^r$ with a norm equivalent to the Sobolev norm as

$$\|f\|_{\mathcal{H}^r}^2 := \sum_{k \leq k_0} \langle f, v_k \rangle^2 + \sum_{k > k_0} \xi_k^r \langle f, v_k \rangle^2,$$

where $k_0$ is any integer satisfying $\xi_k > 0$ for $k > k_0$, thus avoiding the zero eigenvalue (taking $k_0 = 0$ would only result in a semi-norm, not a norm).

Weyl's law for compact Riemannian manifolds (Canzani, 2013, Section 7.8) implies that $\xi_k \asymp k^{2/d}$. Now define the linear operator

$$K := \sum_{k \leq k_0} v_k \otimes v_k + \sum_{k > k_0} \xi_k^{-r} v_k \otimes v_k,$$

where the role of $\tau_k$ is now taken by either 1 or $\xi_k^{-r}$. Since $2r/d > 1$, it implies that $K$ is actually a Hilbert-Schmidt operator acting on $L^2(\mathcal{U})$ (in fact it is trace class) and thus it must also be an integral operator and we can use its kernel as the reproducing kernel of the space. □

According to Theorem 4.4, we have that $h = r/d$ for Sobolev spaces over domains represented as compact Riemannian manifolds (this connection was already known for Euclidean spaces). The minimax rate and phase transition become

$$(nm)^{\frac{-2r}{2r+d}} + n^{-1} \qquad \text{and} \qquad m \asymp n^{d/2r}.$$

We can see the effect of the dimension of the domain on the rates. As we move to higher dimensions the rates get worse, while they improve if the parameters have more derivatives. However, the key point to note is that the rate depends only on the intrinsic dimension of the manifold and not on the dimension of any ambient space. The point where one hits a parametric rate, which is commonly used to distinguish between dense and sparse functional data (Zhang et al.,

2016), is much higher for more complex domains. For example, it is common to assume $r = 2$ derivatives in practice. For one dimensional domains, the phase transition would then occur at $m \asymp n^{1/4}$, which is a relatively easy threshold to meet, while for two dimensions it becomes $n^{1/2}$ and over three it becomes $n^{3/4}$, meaning that one needs nearly as many points per curve as one has subjects, which is a much more stringent threshold.

## 5. Numerical illustrations

In this section we provide a simulation study to numerically explore the estimation error and also provide an application to 3D facial imaging data. Before providing the simulation results and data application, we briefly describe how our estimators are computed.

### *5.1. Computation*

Using the representer theorem one can obtain an exact expression for the estimator. However, this turns out to be very inefficient computationally as it involves solving for $\sum_i m_i$ parameters. Instead, we will approximate the estimator for $\beta_p$ using the first $k_0$ eigenfunctions of $K(u, u')$:

$$\beta_{pk_0}(u) = \sum_{k=1}^{k_0} b_{pk} v_k(u).$$

We provide an exact form for the coefficients $\{b_{pk}\}$. As long as $k_0$ is chosen large enough, then the truncation error will be of a lower order than the convergence rate. If $\beta$ all lie in a $\mathbb{K}$ ball then the truncation error is of the order

$$\|\beta_p - \beta_{pk_0}\|^2 = \sum_{k=k_0+1}^{\infty} b_{pk}^2 = \sum_{k=k_0+1}^{\infty} \tau_k \frac{b_{pk}^2}{\tau_k} \leq \tau_{k_0} \|\beta_p\|_{\mathbb{K}}^2 \asymp k_0^{-2h}.$$

We see that as long as $k_0 \gg n^{1/2h}$ and $k_0 \gg (nm)^{1/(2h+1)}$ then the truncation error will be asymptotically negligible. Of course, in practice, one can take $k_0$ much larger as long as the computational resources allow.

For simplicity, we assume that $m_i \equiv m$, but the general case can be handled by reweighting the $X_{ip}$ and $Y_{ij}$ and using $\bar{m} = \frac{1}{n}\sum_{i=1}^n m_i$ in place of $m$. Let $\mathbf{b} = \{b_{pk}\} \in \mathbb{R}^{P \times k_0}$. The target function is now given by

$$\ell_{nm,\lambda}(\mathbf{b}) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \left( Y_{ij} - \sum_{p=1}^{P} \sum_{k=1}^{k_0} X_{ip} b_{pk} v_k(u_{ij}) \right)^2 + \lambda \sum_{p=1}^{P} \sum_{k=1}^{k_0} \frac{b_{pk}^2}{\tau_k}. \quad (5.1)$$

We will rewrite this expression using vector/matrix notation. First, let $b_v = \mathrm{vec}(\mathbf{b})$, where vec denote stacking the columns into a single vector. Properties

of the vec operation imply that

$$\sum_{k=1}^{k_0} X_{ip} b_{pk} v_k(u_{ij}) = \mathbf{X}_i^\top \mathbf{b} V_{ij} = (V_{ij}^\top \otimes \mathbf{X}_i^\top) b_v$$

where $V_{ij} = (v_1(u_{ij}), \ldots, v_{k_0}(u_{ij}))^\top$. Define

$$Y_v = \text{vec}(\mathbf{Y}),$$

$$\mathbf{A} = \Big\{ (V_{11} \otimes \mathbf{X}_1), (V_{21} \otimes \mathbf{X}_2), \cdots, (V_{n1} \otimes \mathbf{X}_n),$$
$$(V_{12} \otimes \mathbf{X}_1), (V_{22} \otimes \mathbf{X}_2), \cdots, (V_{n2} \otimes \mathbf{X}_n),$$
$$\cdots,$$
$$(V_{1m} \otimes \mathbf{X}_1), (V_{2m} \otimes \mathbf{X}_2), \cdots, (V_{nm} \otimes \mathbf{X}_n) \Big\}^\top \in \mathbb{R}^{(nm)\times(k_0 p)}$$

and let $T$ be a diagonal matrix with its diagonals corresponding to $\{\tau_k\}$, $k = 1, \ldots, k_0$. Then the target function becomes

$$\frac{1}{nm}(Y_v - \mathbf{A}b_v)^\top (Y_v - \mathbf{A}b_v) + b_v^\top (\mathrm{T}^{-1} \otimes \lambda I_P) b_v.$$

The solution can then be expressed as

$$\hat{b}_v = \left((nm)^{-1}\mathbf{A}^\top \mathbf{A} + (\mathrm{T}^{-1} \otimes \lambda I_P)\right)^{-1} (nm)^{-1} \mathbf{A}^\top Y_v.$$

We choose the tuning parameter, $\lambda$, using generalized cross-validation. In the application we allow each $\beta_k$ to have a separate tuning parameter. If $\lambda_k$ is the tuning parameter for $\beta_k$, we can put $\Lambda$ instead of $\lambda I_P$ above where $\Lambda$ is a diagonal matrix with its diagonals corresponding to $\{\lambda_p\}$, $p = 1, \ldots, P$. We cycle several times through each predictor selecting the best value.

### *5.2. Simulation*

In this section, we evaluate the numerical performance of our estimator. We illustrate how the sample size $n$, the number of observations per sample $m$, and different levels of smoothness of the underlying parameters affect the estimation error. We also show that our method is robust against the choice of kernel by examining the estimation error for both the Matérn and Rational Quadratic kernel. Lastly, we highlight the importance of considering the manifold structure of the domain by comparing the estimation error when using the geodesic distance across manifold to the error when using euclidean distance.

The data are generated as

$$Y_{ijl} = \sum_{p=1}^{P} X_{ip} \beta_{lp}(u_{ij}) + \epsilon_{il}(u_{ij}) + \delta_{ijl}$$

where $i = 1, \cdots, n$, $j = 1, \cdots, m$, $l = 1 \cdots, L$. The points $u_{ij}$ are generated uniformly from the domain.

The beta functions are generated using a basis expansion with the basis taken as the eigenfunctions of a Matérn kernel. We use the Matérn kernel as it has parameters that directly controls the smoothness and its resulting RKHS can be tied to a particular Sobelev space (Aronszajn and Smith, 1961, Cho, 2017). The Matérn kernel has the form

$$K(u, w) = \frac{2^{1-\nu}}{\Gamma(\nu)} \frac{\sqrt{2\nu}\|u - w\|^2}{\rho} K_\nu \left( \frac{\sqrt{2\nu}\|u - w\|^2}{\rho} \right)$$

where $K_v$ signifies the modified Bessel function of the second kind of order $\nu$. The smoothing parameter $\nu$ controls the smoothness of the resulting RKHS, and the range parameter $\rho$ scales the distance between $u$ and $s$. Larger $\nu$ would mean that the resulting RKHS will be smoother, and its eigenvalues will decay faster.

We generate the beta function for the simulation setting as

$$\beta^s_{k_s,\nu_s}(u) = \sum_{k=1}^{k_s} v_k(u) + \sum_{k>k_s} \tau_k v_k(u)$$

where $\{v^s_k\}$ are the eigenfunctions of Matérn kernel $K$ with smoothness parameter $\nu_s$ and range parameter $\rho = 1$. The number of leading eigenfunctions for the beta is $k_s$ while the eigenvalues and eigenfunctions of the RKHS are estimated using the algorithm of Pazouki and Schaback (2011). Our error function is also generated using eigenfunctions as $\epsilon_{il}(u_j) = \sum e_{ilk} v_k(u_j)$ with $e_{ilk} \sim N(0, \tau^4_k)$ for $L = 1$. For the case of $L = 2$, we make $\epsilon_{i1}(u)$ and $\epsilon_{i2}(u)$ correlated by generating their coefficients for $v_1(u)$ with $\left(\begin{smallmatrix} e_{i11} \\ e_{i21} \end{smallmatrix}\right) \sim N\left(\left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right), \tau^4_1 \left(\begin{smallmatrix} 1 & 0.3 \\ 0.3 & 1 \end{smallmatrix}\right)\right)$. The measurement error $\delta_{ij}$ is generated from $N(0, 0.1)$.

We consider several domains, $\mathcal{U}$, including a line, a plane, a parabola, and a swiss roll. Each setting is different in terms of the smoothness of RKHS where the beta lies ($\nu_s$) and the number of leading eigenfunctions ($k_s$). The resulting $\beta^s_{k_s,\nu_s}$ functions are shown in Figures 1, 2, and 3. We conducted the combination of setting as the following. Here, $d$ represents the dimension of $\mathcal{U}$, $L$ represents the dimension of the response, and $P$ represents the number of the predictors.

1. A line $\mathcal{U}$ $(d = 1)$

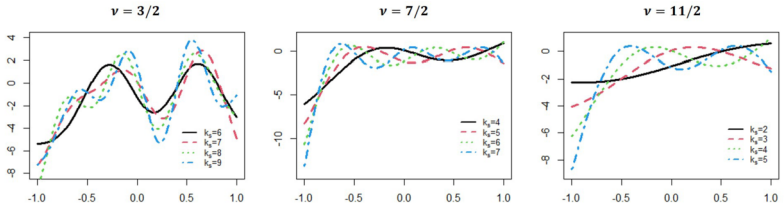


FIG 1. *The beta functions generated for the line domain.*

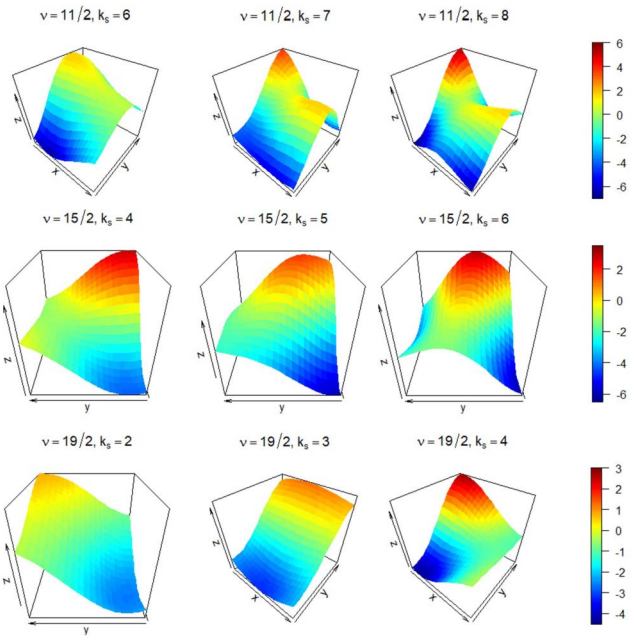


FIG 2. *The beta functions generated for the plane domain.*

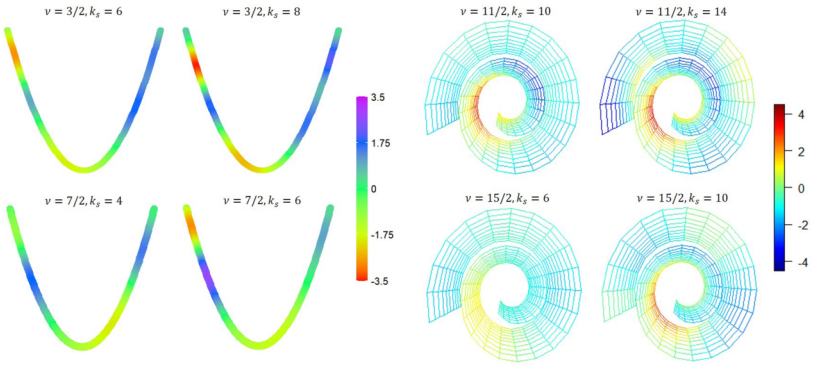


FIG 3. *The beta functions generated for the parabola and swiss roll domain.*

(a) $L = 1$, $P = 2$: $X_1 = 1$ (intercept) and $X_2 \sim N(1, 1)$
   i. $\nu_s = 3/2$: $\beta_1(u) = \beta^s_{8,3/2}(u)$, $\beta_2(u) = \beta^s_{7,3/2}(u)$
   ii. $\nu_s = 7/2$: $\beta_1(u) = \beta^s_{5,7/2}(u)$, $\beta_2(u) = \beta^s_{6,7/2}(u)$
   iii. $\nu_s = 11/2$: $\beta_1(u) = \beta^s_{4,11/2}(u)$, $\beta_2(u) = \beta^s_{3,11/2}(u)$

(b) $L = 2$, $P = 2$: $X_1 = 1$ (intercept) and $X_2 \sim N(1, 1)$
   i. $\nu_s = 3/2$: $\beta_{11}(u) = \beta^s_{8,3/2}(u)$, $\beta_{21}(u) = \beta^s_{6,3/2}(u)$, $\beta_{12}(u) =$

$\beta^s_{9,3/2}(u)$,
$\beta_{22}(u) = \beta^s_{7,3/2}(u)$

ii. $\nu_s = 7/2$: $\beta_{11}(u) = \beta^s_{7,7/2}(u)$, $\beta_{21}(u) = \beta^s_{5,7/2}(u)$, $\beta_{12}(u) = \beta^s_{6,7/2}(u)$,
$\beta_{22}(u) = \beta^s_{4,7/2}(u)$

iii. $\nu_s = 11/2$: $\beta_{11}(u) = \beta^s_{5,11/2}(u)$, $\beta_{21}(u) = \beta^s_{3,11/2}(u)$, $\beta_{12}(u) = \beta^s_{4,11/2}(u)$, $\beta_{22}(u) = \beta^s_{2,11/2}(u)$

(c) $L = 1$, $P = 3$: $X_1 = 1$ (intercept), $X_2 \sim N(1,1)$, and $X_3 \sim N(1,1)$

i. $\nu_s = 3/2$: $\beta_1(u) = \beta^s_{8,3/2}(u)$, $\beta_2(u) = \beta^s_{7,3/2}(u)$, $\beta_3(u) = \beta^s_{6,3/2}(u)$

ii. $\nu_s = 7/2$: $\beta_1(u) = \beta^s_{5,7/2}(u)$, $\beta_2(u) = \beta^s_{6,7/2}(u)$, $\beta_3(u) = \beta^s_{4,7/2}(u)$

iii. $\nu_s = 11/2$: $\beta_1(u) = \beta^s_{4,11/2}(u)$, $\beta_2(u) = \beta^s_{3,11/2}(u)$, $\beta_3(u) = \beta^s_{2,11/2}(u)$

(d) $L = 1$, $P = 3$: $X_1 \sim N(1,1)$, $X_2 \sim N(1,1)$, and $X_3 = X_1 \times X_2$ (interaction)

i. $\nu_s = 3/2$: $\beta_1(u) = \beta^s_{8,3/2}(u)$, $\beta_2(u) = \beta^s_{7,3/2}(u)$, $\beta_3(u) = \beta^s_{6,3/2}(u)$

ii. $\nu_s = 7/2$: $\beta_1(u) = \beta^s_{5,7/2}(u)$, $\beta_2(u) = \beta^s_{6,7/2}(u)$, $\beta_3(u) = \beta^s_{4,7/2}(u)$

iii. $\nu_s = 11/2$: $\beta_1(u) = \beta^s_{4,11/2}(u)$, $\beta_2(u) = \beta^s_{3,11/2}(u)$, $\beta_3(u) = \beta^s_{2,11/2}(u)$

2. A plane $\mathcal{U}$ ($d = 2$)

(a) $L = 1$, $P = 2$: $X_1 = 1$ (intercept) and $X_2 \sim N(1,1)$

i. $\nu_s = 11/2$: $\beta_1(u) = \beta^s_{8,11/2}(u)$, $\beta_2(u) = \beta^s_{7,11/2}(u)$

ii. $\nu_s = 15/2$: $\beta_1(u) = \beta^s_{5,15/2}(u)$, $\beta_2(u) = \beta^s_{6,15/2}(u)$

iii. $\nu_s = 19/2$: $\beta_1(u) = \beta^s_{4,19/2}(u)$, $\beta_2(u) = \beta^s_{3,19/2}(u)$

(b) $L = 1$, $P = 3$: $X_1 = 1$ (intercept), $X_2 \sim N(1,1)$, and $X_3 \sim N(1,1)$

i. $\nu_s = 11/2$: $\beta_1(u) = \beta^s_{8,11/2}(u)$, $\beta_2(u) = \beta^s_{7,11/2}(u)$, $\beta_3(u) = \beta^s_{6,11/2}(u)$

ii. $\nu_s = 15/2$: $\beta_1(u) = \beta^s_{5,15/2}(u)$, $\beta_2(u) = \beta^s_{6,15/2}(u)$, $\beta_3(u) = \beta^s_{4,15/2}(u)$

iii. $\nu_s = 19/2$: $\beta_1(u) = \beta^s_{4,19/2}(u)$, $\beta_2(u) = \beta^s_{3,19/2}(u)$, $\beta_3(u) = \beta^s_{2,19/2}(u)$

3. A parabola $\mathcal{U}$ ($d = 1$)

(a) $L = 1$, $P = 2$: $X_1 = 1$ (intercept) and $X_2 \sim N(1,1)$

i. $\nu_s = 3/2$: $\beta_1(u) = \beta^s_{6,3/2}$, $\beta_2(u) = \beta^s_{8,3/2}$

ii. $\nu_s = 7/2$: $\beta_1(u) = \beta^s_{4,7/2}$, $\beta_2(u) = \beta^s_{6,7/2}$

4. A swiss roll $\mathcal{U}$ ($d = 2$)

(a) $L = 1$, $P = 2$: $X_1 = 1$ (intercept) and $X_2 \sim N(1,1)$

i. $\nu_s = 11/2$: $\beta_1(u) = \beta^s_{10,11/2}$, $\beta_2(u) = \beta^s_{14,11/2}$

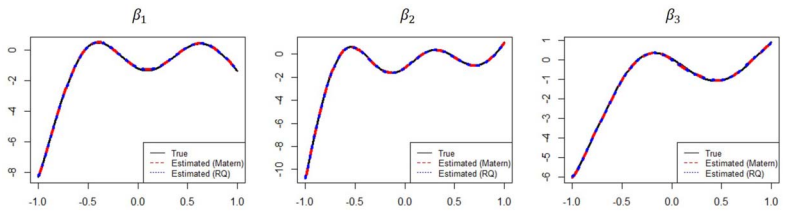


FIG 4. *Examples of estimated beta functions for swiss roll domain for setting 1(d)ii with $n = 50$, $m = 50$, and simulation run number of 318. The estimated betas with Matérn kernel and the estimated betas with Rational Quadratic kernel are almost the same as the true betas.*

$$\text{ii. } \nu_s = 15/2\text{: } \beta_1(u) = \beta^s_{6,15/2},\ \beta_2(u) = \beta^s_{10,15/2}$$

For each setting of the line $\mathcal{U}$ and the plane $\mathcal{U}$, we tried all combinations of $n = 10, 20, 30, 40, 50$ and $m = 10, 20, 30, 40, 50$ and for each setting of the parabola $\mathcal{U}$ and the swiss roll $\mathcal{U}$, we tried all combinations of $n = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$ and $m = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$. And we ran 1000 repetitions of each scenario and each combination. For estimation, we use the Matérn kernel and Rational Quadratic kernel. Rational Quadratic kernel has the form of

$$K(u, w) = \left(1 + \frac{\|u - w\|^2}{\alpha l^2}\right)^{-\alpha}.$$

The choice of parameters, such as the smoothness parameter $\nu$ and the range parameter $\rho$ for the Matérn RKHS, and the $\alpha$ and $l$ of Rational Quadratic kernel are done using the generalized cross validation (GCV). The choice of $\lambda$ in (5.1) is also done through GCV. Examples of estimated functional coefficients are presented in Figure 4 (line) and Figure 5 (swiss roll).

We report the estimation error as $R^s_n = \sum_{p=1}^{P} \sum_{l=1}^{L} \|\hat{\beta}^s_{lp} - \beta^s_{lp}\|^2$ for each run and the mean estimation error of 1000 simulation runs for each $n$ and $m$ are shown in Figure 6, Figure 7, and Figure 8. For ease of exposition, we only show the results for the line and swiss roll, but the results for the plane and parabola can be found in the appendix. For each $5 \times 5$ heatmap in Figures 6 and 7, the bottom leftmost presents the mean estimation error for $n = 10$ and $m = 10$; as one moves right, $n$ increases, and as one moves up, $m$ increases. Therefore, the top rightmost presents the estimation error for $n = 50$ and $m = 50$. For each 10 by 10 heatmap in Figure 8, the bottom leftmost represents $n = 10$ and $m = 10$ whereas the top rightmost represents $n = 100$ and $m = 100$.

**Discussion**: For all simulation settings, the mean squared estimation error decreases as $n$ and $m$ increase, as indicated by the darker blue color in each of the heatmaps in Figures 6, 7, and 8. There are a few anomalies presented; for example, the top left plot in Figure 6 presents the estimation errors for the setting 1(a)i (line $\mathcal{U}$ with $P = 2$, $L = 1$, and $\nu = 3/2$). The top right corner shows a little higher mean estimation errors, such as 0.081, 0.098, and 0.059, but these are due to some outliers. The median estimation errors for the same

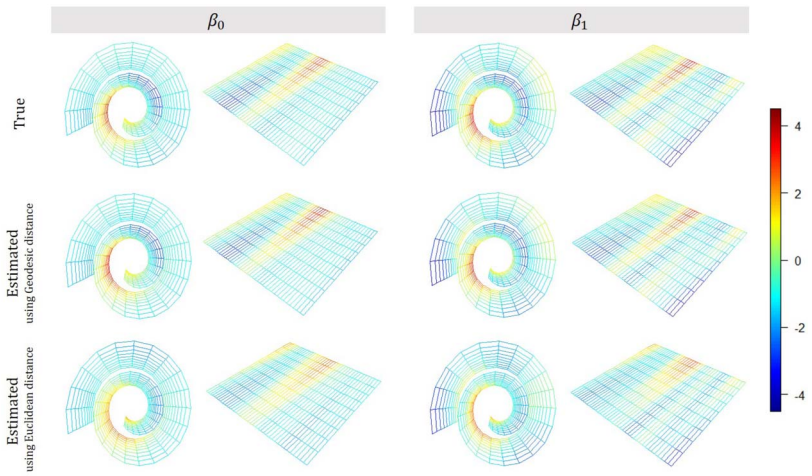


FIG 5. *Examples of estimated beta functions for swiss roll domain for setting 4(a)i ($\nu = 11/2$) with $n = 50$, $m = 50$, and simulation run number of 179. Both rolled version and unrolled version are presented for comparison. The estimated betas using Geodesic distance (middle row) are almost same as the true beta, but the estimated betas using Euclidean (bottom row) are different from the true beta.*

cells were 0.015, 0.013, and 0.012.

We can also see the effect of the *phase-transition* as $n^{-1}$ becomes the dominating component of the convergence rate for larger values of $m$, as the mean squared estimation error does not change much when $m$ is sufficiently large but continues to decrease with increasing $n$. For example, when we check the top middle plot in Figure 6 which presents the setting 1(a)ii, the estimation error is 0.095 for $n = 20$ and $m = 10$, and this drops to 0.05 for $n = 20$ and $m = 20$ and stays as 0.045, 0.044, and 0.042 as $m$ increases to 30, 40, and 50. However, when we increase $n$, the estimation error keeps decreasing.

For the same $n$ and $m$, the mean squared estimation error decreases as $\nu$ increases, or the true beta lies in the smoother RKHS. For example, for $n = 10$ and $m = 10$, the estimation error for $\nu = 3/2$ is 15.688, that for $\nu = 5/2$ is 4.238, and that for $\nu = 7/2$ is 0.211. But it has also been observed that with high enough $n$ and $m$, increasing $\nu$ does not really change the estimation error, likely since $n^{-1}$ becomes the dominating component of the convergence rate.

When we compare $L = 1$ cases and $L = 2$ cases in Figure 6, the mean estimation error is larger with $L = 2$ as expected. When we compare $P = 2$ cases in Figure 6 to $P = 3$ cases in Figure 7, the mean estimation error is higher with $P = 3$, also as expected. The performance of our estimator stays the same when interaction term is included as in the bottom six heatmaps in Figure 7 which are for settings 1(d)i, 1(d)ii, and 1(d)iii.

When we compare the estimation using Matérn kernel and the estimation using Rational Quadratic kernel, the latter shows a slightly higher estimation error, but they are very similar. For example, the ratio of estimation error with

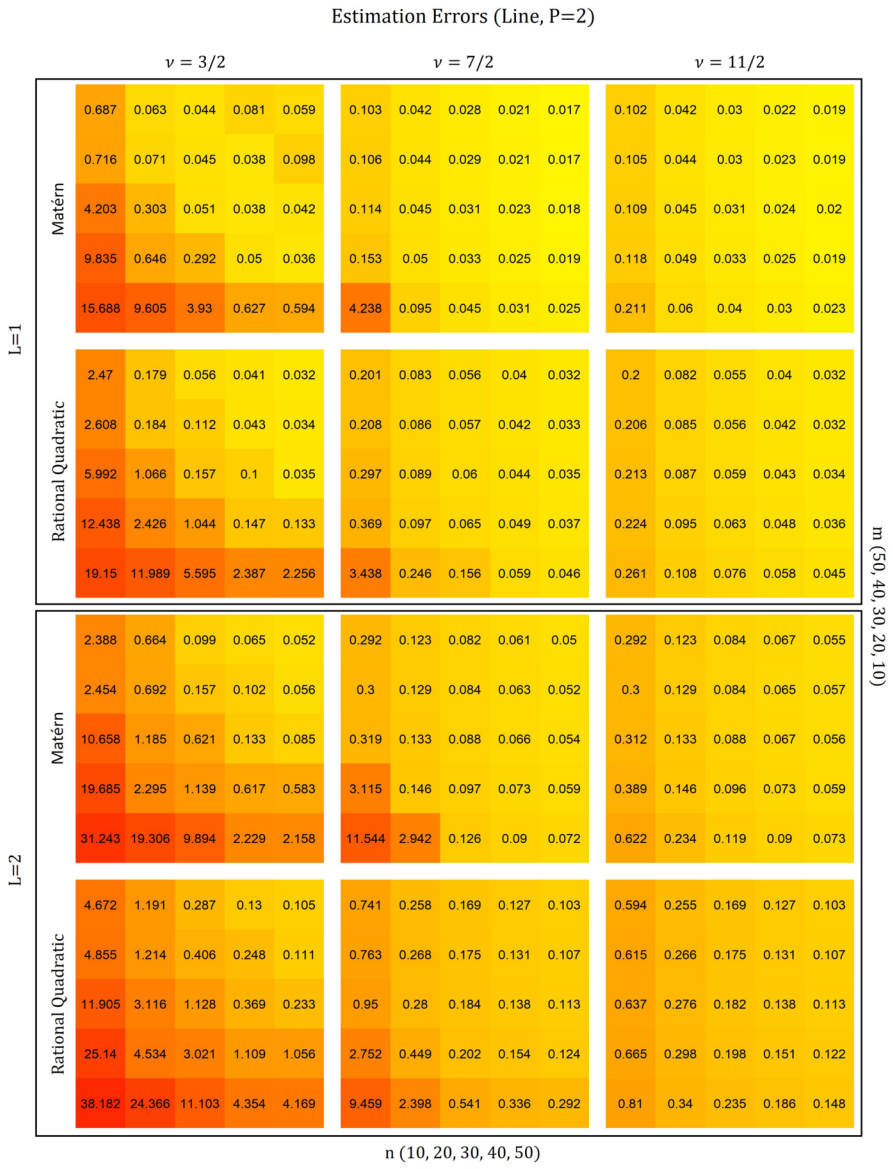


FIG 6. *The effects of $n$ and $m$ on the mean squared estimation errors for $L = 1$ and $L = 2$ case where domain is line and $P = 2$. For each heatmap, the bottom leftmost cell presents the mean estimation error for $n = 10$, $m = 10$, and as it moves towards right, $n$ increases, whereas as it moves towards the top, $m$ increases. Therefore, the top rightmost cell presents the estimation error for $n = 50$, $m = 50$.*

Matérn divided by the estimation error with Rational Quadratic has mean of 0.84 and median of 0.99 for the cases presented in Figure 8. Also Figure 4

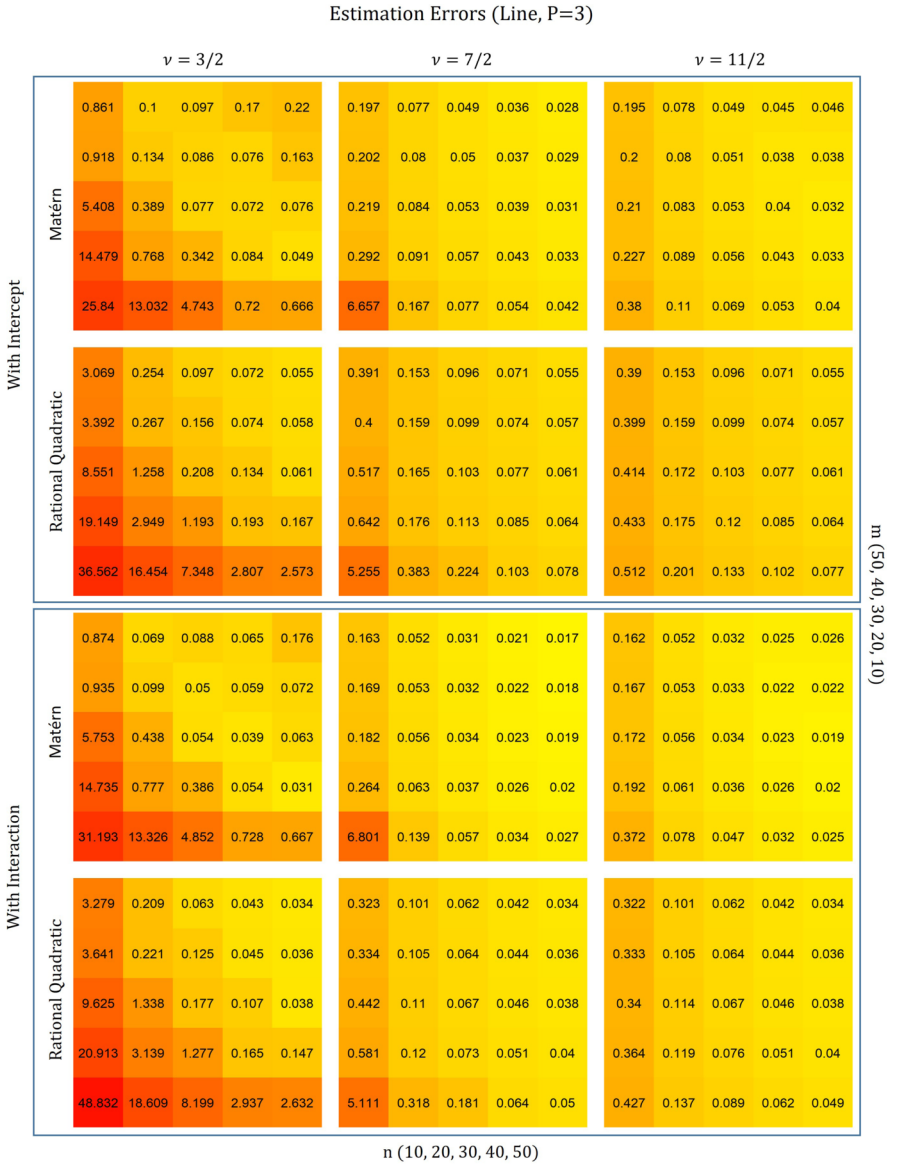


FIG 7. *The effects of $n$ and $m$ on the mean squared estimation errors for line domain with $P = 2$. On the top it shows the setting 1(c) with an intercept, and on bottom it shows the setting 1(d) with an interaction term.*

shows that the estimated beta with Matérn kernel and the estimated beta with Rational Quadratic kernel are almost the same, and both are very close to the true beta.

When we compare the estimation error using geodesic distance to the estima-

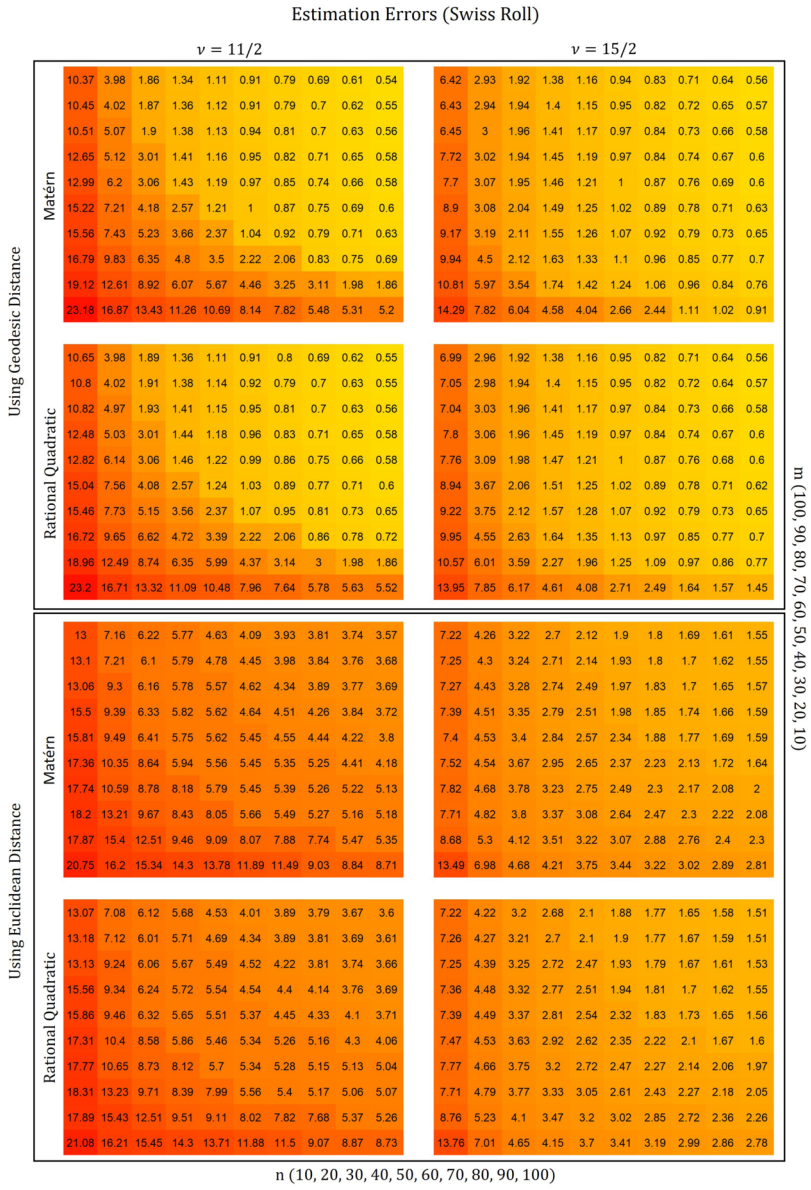


FIG 8. *The effects of n and m on the mean squared estimation errors for swiss roll domain.*

tion error using Euclidean distance, it is clear that the estimation error found with Euclidean distance is much larger than that with geodesic distance. On average, the estimation error with geodesic distance was 42% lower and at maximum it was 88% lower for cases presented in Figure 8. Also Figure 5 shows that the estimated beta using Geodesic distance is almost the same as the true beta,

whereas the estimated beta using Euclidean distance is different from the true beta. This emphasizes the importance of considering the manifold structure and using geodesic distance when we have manifold domain.

### *5.3. 3D facial data*

We apply our optimal estimator to the facial data collected through the Penn State ADAPT study (Claes et al., 2014a,b). Following the framework of Kang et al. (2017) (who based their models on felsplines and FPCA), we fit a manifold-on-scalar regression model with the dependent/outcome variable being a 3D human facial face parametrized by a two-dimensional manifold $\mathcal{U}$ representing a common template face (we use the average face), and the independent / explanatory variables as sex, age, height, weight, and genetic ancestry. Genetic ancestry is measured as the proportions from particular ethnic backgrounds: Northern Europe, Southern Europe, East Asia, South Asia, Native America, and West Africa. We also include interactions between sex and age, age and weight, and height and weight.

The faces are densely measured with 7150 points in x, y, and z coordinates, so $Y_{ijl}$ in (1.1) will be the measurement of the $j$-th point of the $i$-th person's face in the $l$-th coordinate. The sample size is $n = 3287$, with $m = 7150$ and $L = 3$. Since the template face, $\mathcal{U}$, is two-dimensional manifold, $d = 2$. There are in total $P = 13$ predictors including the intercept term. Prior to model fitting all faces are scaled and aligned using generalized procrustes analysis. The computation follows section 5.1, and the choices of $\lambda$, $\nu$, and $\rho$ are done through GCV.

Four of the resulting $\hat{\beta}_p$'s, which are 3-dimensional functional objects, are shown in Figure 9. We only present the key predictors here, but other $\hat{\beta}$'s are presented in the appendix. The middle plot in Figure 9 is the predicted face of a Northern European male, aged 30, with height of 170cm and weight of 70kg. In each corner we repeat the prediction, but with one covariate value changed. On the top left is the predicted face of a Northern European female with the same age, height, and weight. The red and blue plot in between is the visualized estimated beta for sex. Red means there is a shift of the face outward, and blue means inward. From male to female, the red on the cheek and the blue on the chin show that the face becomes a bit rounder, and the red on the eyelids and the blue on the eyebrow give less prominent eyebrows and rounder eyes. Also, the slight hint of red around the nostrils show that female would also have a bit rounder nose.

On the top right is the predicted face when changing the age from 30 to 60 years old. The red and blue plot in between is again the visualized estimated beta for age. The red in the cheeks and jawline and the blue in between them show that the skin hangs more loosely on the cheek and jawline area, which (unfortunately) is a common aging effect. Another noticeable effect is on the eyes; the loose skin on eyelids and the bags under eyes are also well-known aging effects, and this is captured in the beta plot with the red on the eyelids

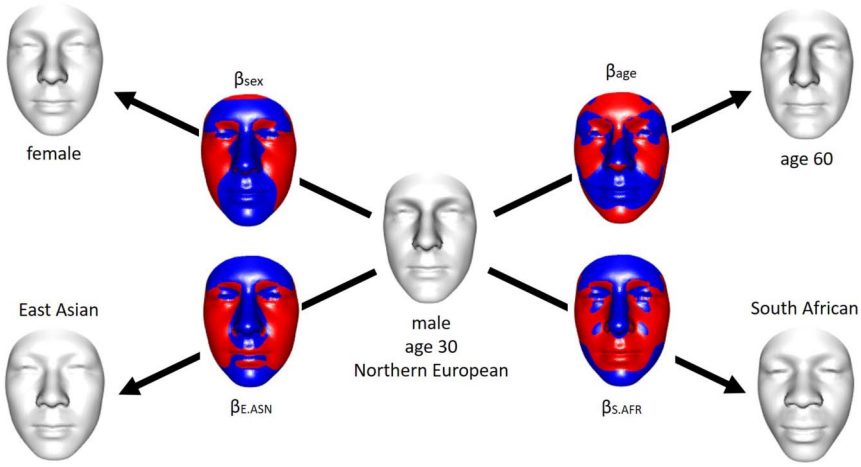


FIG 9. *The middle grey plot is the predicted face of a Northern European male with age of 30, height of 170cm, and weight of 70kg. The four grey plots on the far sides are the predicted faces with one predictor change from the middle plot. The red and blue plots are the visualization of the effects of the corresponding estimated betas where the red means outward effect and the blue means inward effect.*

and under the eye, and with the blue in the middle and on the sides of the eyes.

On the bottom left is the corresponding predicted face for East Asian ancestry, and the corresponding colored plot shows how the predicted faces differ between Northern Europeans and East Asians. The red on the cheek and the blue on the chin shows that the predicted East Asian has a rounder face, and the blue on the nose with a little red on the sides mean that the predicted East Asian has a less prominent and slightly rounder nose than Northern European. Also, the predicted East Asian has less prominent eyebrows and forehead as the blue on those areas shows, and he has rounder eyes.

On the bottom right is the predicted face for South African ancestry. The plots indicate that the nose of the predicted South African is flatter and wider with the blue in the middle of nose and the red on the sides of nose. There seems to be minor tear-through nasojugal grooves under the eyes and the nasolabial folds below the nose in the predicted face of a South African, and these lines are captured with the blue dots on the cheeks.

## 6. Conclusions

In this work we have presented new results concerning minimax rates for a function-on-scalar regression when the domain of the functions is more complex than just an interval. Assuming the parameters reside in an RKHS results in the rates being closely tied to the decay of the eigenvalues. However, the rates in such cases, and thus the difficulty of the problem, can be somewhat hidden behind the eigenvalues. To add clarity to our results, we extend well known

connections between RKHS and Sobolev spaces to the case where the domain of the functions are compact Riemannian manifolds.

A great deal of biomedical imaging data is being collected and analyzed in scientific studies. As our technologies progress, such statistical methods will become increasingly important. This is especially critical if statistical tools are to keep pace with more "black box" machine learning methods. Indeed, though the data is complicated, a major selling point of our methodology (and most statistical methods) is the ability to provide clear interpretations for the effects in our model, which scientists and practitioners will find useful.

We provided a practical strategy for implementing our methods via a basis representation based on the RKHS kernel being used, which avoids some of the large matrix inversion problems inherent in using the representer theorem. This approach scales nicely and provides a flexible tool that can be applied in a variety of settings so long as the RKHS kernel can be defined and computed. However, we don't view this estimator as definitive and would be excited to see what insights other researchers have when choosing kernels and modelling strategies for different applications.

## Appendix A: Proof of lower bound

In the following we give an adaptation of the proof in Cai and Yuan (2011) for the case of general RKHS. Interestingly, the lower bound is only tight if the harmonic and arithmetic mean are asymptotically equivalent, that is, they grow at the same order with $n$. This stems from their upper bound being in terms of the harmonic mean, but the arguments for the lower bound lead to the arithmetic mean (if one does not assume the $m_i$ are identical). We also provide some extra details for the interested reader. To prove the lower bound result, we will employ Fano's lemma and construct an example that achieves the worst case rate.

Recall that for lower bounds, we need only find one model $M \in \mathcal{M}$, that achieves the desired rate. We can thus make any assumptions we like as long as it remains a valid model. Therefore, assume that the $\epsilon_{i\ell}$ and $\delta_{ij\ell}$ are iid across $i, j$ and $\ell$. We do not assume that $X_i \equiv 1$, since nowhere did we say that the intercept is always included in the model. The parameter, $\boldsymbol{\beta} = \{\beta_{\ell,p}\}$, we view as a vector of $LP$ functions, and thus an element of $\mathbb{K}^{LP}$. Let $B_{\mathbb{K}}$ be the unit ball in $\mathbb{K}$. In this case the model is given by

$$Y_{ij\ell} = \mathbf{X}_i^\top \boldsymbol{\beta}_\ell(u_{ij}) + \varepsilon_{i\ell}(u_{ij}) + \delta_{ij\ell},$$

and we assume that each $\boldsymbol{\beta}_{\ell p} \in B_{\mathbb{K}}$. Assume that the distribution of the observed locations $u_{ij}$ has a uniform density with respect to the base measure $\mu$. Assume that $\varepsilon_{i\ell}$ are iid mean zero Gaussian processes with covariance function, $C$, and that the $\delta_{ij\ell}$ are iid mean zero normals with variance 1. Consider $M$ different parameters that we will define more explicitly later, $\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^M \in \mathbb{K}^{LP}$ and their induced probability measures $P_1, \ldots, P_M$ for the $\{Y_{ij\ell}\}$. Fano's lemma tells us the following.

**Lemma A.1** (Fano's Lemma)**.** *Let $P_1, \dots, P_M$ be probability measures over a measurable space $(\Omega, \mathcal{F})$, such that*

$$KL(P_k||P_{k'}) \leq \alpha, \qquad k \neq k',$$

*then for any test function $\psi : \Omega \to \{1, \dots, M\}$ we have*

$$P_k(\psi = k) \leq \frac{\alpha + \log 2}{\log(M-1)} \qquad or \qquad P_k(\psi \neq k) \geq 1 - \frac{\alpha + \log 2}{\log(M-1)}.$$

In other words, Fano's lemma gives us an upper bound on the estimation accuracy for any possible test we could construct to select the true $\boldsymbol{\beta}$ from among the $\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^M$. Any estimator, $\hat{\boldsymbol{\beta}}$, we could construct in this setting would be equivalent to choosing one of the $\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^M$. Thus, in this case the estimation error must be at least

$$\begin{aligned} \mathrm{E}_{P_k} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^k\|^2 &\geq P_k(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^k) \min_{k,k'} \|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k'}\|^2 \\ &\geq \left(1 - \frac{\alpha + \log 2}{\log(M-1)}\right) \min_{k,k'} \|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k'}\|^2 \end{aligned}$$

for any estimator. So, applying Fano's lemma becomes a matter of selecting $\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^M$ that are well separated while properly balancing the KL divergence.

To compute the KL divergence, we first condition on the domain points, $u_{ij}$, as then the $Y_{i\ell j}$ are all Gaussian. We can then take an expectation with respect to the $u_{ij}$ to complete the computation. Recall that between two Gaussian random vectors, $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, it is given by $(1/2)(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$. Each $P_k$, which is the distribution of $\{Y_{i\ell j} : i = 1, \dots, n;\ \ell = 1, \dots, L;\ j = 1, \dots, m_i\}$, is composed of $nL$ independent Gaussian measures (conditioned on the $u_{ij}$), over which the KL divergence is additive. Let $\mathbf{u}_i \sim (u_{i1}, \dots, u_{im_i})$ and $\Sigma(\mathbf{u}_i) := \{C(u_{ij}, u_{ij'}) + 1_{j=j'}\}$, then $P_i$ is composed of $n$ blocks of size $L$ that all of the same conditional covariance $\Sigma(\mathbf{u}_i)$. Since $\Sigma(\mathbf{u}_i)$ is the sum of two positive definite matrices, one being the identity, we have that $\Sigma(\mathbf{u}_i) \succeq \mathbf{I}_{m_i}$ as positive definite matrices. This also implies that $\Sigma(\mathbf{u}_i)^{-1} \preceq \mathbf{I}_{m_i}$. Define the $P \times m_k$ matrix $\beta_\ell^k(\mathbf{u}_i) = \{\beta_{\ell p}^k(u_{ij}) : p = 1, \dots, P;\ j = 1, \dots, m_k\}$, then $P_{ik}$, conditioned on $\mathbf{u}_k$, is Gaussian with mean $\beta_\ell^i(\mathbf{u}_k)^\top \mathbf{X}_i$ and covariance $\Sigma(\mathbf{u}_k)$. To bound the KL divergence we first need to compute, for each $\ell = 1, \dots, L$ and $i = 1, \dots, n$

$$\begin{aligned} &\mathrm{E}[(\mathbf{X}_i^\top \beta_\ell^k(\mathbf{u}_i) - \mathbf{X}_i^\top \beta_\ell^{k'}(\mathbf{u}_i))\Sigma(\mathbf{u}_i)^{-1}(\mathbf{X}_i^\top \beta_\ell^k(\mathbf{u}_i) - \mathbf{X}_i^\top \beta_\ell^{k'}(\mathbf{u}_i))^\top] \\ &\leq \mathrm{E}[\mathbf{X}_i^\top (\beta_\ell^k(\mathbf{u}_i) - \beta_\ell^{k'}(\mathbf{u}_i))\mathbf{I}_{m_i}(\beta_\ell^k(\mathbf{u}_i) - \beta_\ell^{k'}(\mathbf{u}_i))^\top \mathbf{X}_i] \\ &= \mathrm{E}\left[\left|(\beta_\ell^k(\mathbf{u}_i) - \beta_\ell^{k'}(\mathbf{u}_i))^\top \mathbf{X}_i\right|^2\right] \\ &= \sum_{r=1}^{m_i} \mathrm{E}[(\mathbf{X}_i^\top \beta_\ell^k(u_{kr}) - \mathbf{X}_i^\top \beta_\ell^{k'}(u_{kr}))^2] = m_i \|\mathbf{X}_i^\top \beta_\ell^k - \mathbf{X}_i^\top \beta_\ell^{k'}\|^2, \end{aligned}$$

since $u_{kj}$ is uniform over $\mathcal{U}$. Since the KL diveregence, conditioned on the $\mathbf{u}_i$ is addittive, we simply need to added up the above quantity over $i$ and $\ell$, then

divide by two to get the following bound:

$$KL(P_k||P_{k'}) \leq \frac{1}{2}\sum_{i=1}^{n} m_i \sum_{\ell=1}^{L} \|\mathbf{X}_i^\top \boldsymbol{\beta}_\ell^k - \mathbf{X}_i^\top \boldsymbol{\beta}_\ell^{k'}\|^2$$
$$\leq \frac{\zeta n m_a \max_{kk'} \|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k'}\|^2}{2}$$

where $m_a$ is the arithmetic mean and from Assumption 2.1, $\zeta$ bounds $|\mathbf{X}_k|^2$ for all $k$. Our estimation error is then bounded from below by

$$\left(1 - \frac{(\zeta^2/2) n m_a \max_{kk'} \|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k'}\|^2 + \log(2)}{\log(M-1)}\right) \min_{kk'} \|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k'}\|^2.$$

We want to make this error as large as possible (since that would produce the tightest lower bound), under the constraint that each of the $LP$ parameters lie in the unit ball in $\mathbb{K}$, i.e. $\boldsymbol{\beta}_{\ell p}^k \in B_{\mathbb{K}}$, for $\ell = 1, \ldots, L$ amd $p = 1, \ldots, P$. To construct a viable sequence, we consider the Varshamov-Gilbert bound (Varshamov, 1957, Duchi, 2016).

**Lemma A.2** (Varshamov-Gilbert)**.** *For $N \geq 1$ there exists at least $M = \exp(NLP/8)$ $NLP$-dimensional vectors, $b_1, \ldots, b_M$, with entries $b_{kj} \in \{0, 1\}$ such that*

$$\sum_{j=1}^{NLP} 1\{b_{kj} \neq b_{k'j}\} \geq NLP/4, \qquad \text{for } k \neq k'.$$

This is a commonly used lemma for constructing collections of parameters for minimax results as they take a very simple form. We rearrange the vectors $b_k$ into arrays $\mathbf{b}^k$ of dimension $N \times L \times P$. We can use these sequences to construct elements of $L^2(\mathcal{U})$ in the $v_k$ basis. Define

$$\beta_{\ell,p}^k := N^{-1/2} \sum_{j=N+1}^{2N} \tau_i^{1/2} b_{j-N,\ell,p}^k v_i \qquad k = 1, \ldots, M.$$

Recall that for any norm we have $\|\boldsymbol{\beta}^k\|^2 = \sum_{\ell,p} \|\beta_{\ell p}^k\|^2$. Then we have the following properties, for $k \neq k'$, through a direct verification

$$\|\boldsymbol{\beta}^k\|_{\mathbb{K}}^2 \leq LP, \qquad \|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k'}\|^2 \geq LP\tau_{2N}/4, \qquad \|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k'}\|^2 \leq LP\tau_N.$$

Using this sequence, the lower bound becomes

$$\left(1 - \frac{(\zeta^2/2) n m_a LP\tau_N + \log(2)}{NLP/8}\right) LP\tau_{2N}/4$$
$$\asymp \left(1 - \frac{4\zeta^2 n m_a LPN^{-2h} + 8\log(2)}{NLP}\right) LPN^{-2h}.$$

Taking $N = (8\zeta^2 n m_a)^{1/(1+2h)}$, which implies $N \to \infty$, would produce

$$\left(\frac{1}{2} - \frac{8\log(2)}{NLP}\right) LP(2N)^{-2h} \asymp LP(nm_a)^{-2h/(1+2h)},$$

which is the desired bound as long as $m_a \asymp m$. This bound (as we will see) matches the upper bound in the case where $m \asymp m_a$ and $m \ll n^{1/2h}$ or is of the same order, giving a tight rate. However, in the case where $m \gg n^{1/2h}$, then the bound is loose.

To obtain a bound that works when $m \gg n^{1/2h}$ we can make the problem even simpler. Assume that $\beta_{\ell p}^k(u) \equiv a_{\ell p}^k \in \mathbb{R}$ and that $C(u, u') \equiv 1$, meaning that there are no dynamics in time (one just has a repeated measures problem). Let $\mathbf{a}_\ell^k := \{a_{\ell p}^k\}$ be the $P$-dimensional vector of slope coefficients. A simple verification shows that the vector of all ones is an eigenvector of $\Sigma(\mathbf{u}_i) = \mathbf{I}_{m_i} + \mathbf{1}_{m_i}\mathbf{1}_{m_i}^\top$ with eigenvalue $m_i + 1$ and $\mathbf{1}_{m_i}$ is an $m_i$ dimensional vector of all ones. Furthermore, we have the simplification $\beta_\ell^k(\mathbf{u}_i) = \mathbf{a}_{lp}^k \mathbf{1}_{m_i}^\top$. This implies that the KL divergence is now bounded by

$$\begin{aligned}
&KL(P_k||P_{k'}) \\
&= \frac{1}{2}\sum_{i=1}^n \sum_{\ell=1}^L (\mathbf{X}_i^\top \mathbf{a}_\ell^k \mathbf{1}_{m_i}^\top - \mathbf{X}_i^\top \mathbf{a}_\ell^{k'} \mathbf{1}_{m_i}^\top)\Sigma(\mathbf{u}_i)^{-1}(\mathbf{X}_i^\top \mathbf{a}_\ell^k \mathbf{1}_{m_i}^\top - \mathbf{X}_i^\top \mathbf{a}_\ell^{k'} \mathbf{1}_{m_i}^\top)^\top \\
&= \frac{1}{2}\sum_{\ell=1}^L \sum_{i=1}^n (m_i+1)^{-1}(\mathbf{X}_i^\top \mathbf{a}_\ell^k \mathbf{1}_{m_i}^\top - \mathbf{X}_i^\top \mathbf{a}_\ell^{k'} \mathbf{1}_{m_i}^\top)(\mathbf{X}_i^\top \mathbf{a}_\ell^k \mathbf{1}_{m_i}^\top - \mathbf{X}_i^\top \mathbf{a}_\ell^{k'} \mathbf{1}_{m_i}^\top)^\top \\
&= \frac{1}{2}\sum_{\ell=1}^L \sum_{i=1}^n \frac{m_i}{m_i+1}(\mathbf{X}_i^\top \mathbf{a}_\ell^k - \mathbf{X}_i^\top \mathbf{a}_\ell^{k'})^2 \\
&\leq \sum_{\ell=1}^L \frac{n\zeta^2 |\mathbf{a}_\ell^k - \mathbf{a}_\ell^{k'}|^2}{2} = \frac{n\zeta^2 |\mathbf{a}^k - \mathbf{a}^{k'}|^2}{2}.
\end{aligned}$$

To construct our sequence, recall that the unit ball in $\mathbb{R}^{L\times P}$ has a $1/2$ packing number greater than $2^{LP}$, meaning, we can find at least $2^{LP}$ matrices, $\mathbf{b}^1, \mathbf{b}^2, \ldots,$ within the unit ball that are at least $1/2$ units apart (Duchi, 2016). Set $\mathbf{a}^k = \delta \mathbf{b}^k$, which we will specify in a moment. Then clearly $|\mathbf{a}^k - \mathbf{a}^{k'}| \leq 2\delta$ and $|\mathbf{a}^k - \mathbf{a}^{k'}| \geq \delta/2$, for $k \neq k'$. This implies a lower bound of

$$\left(1 - \frac{2\zeta n\delta^2 + \log(2)}{LP\log(2)}\right)\delta^2.$$

Now, simply set $\delta^2 \asymp LP/n$ to obtain the desired result.

## Appendix B: Proof of upper bound

Since each coordinate of the response can be estimated separately, we will assume wlog that $L = 1$ in our proof. We also assume, wlog, that $u_{ij}$ have a

density identically equal to 1, meaning their law is given by $\mu$. We assume the kernel $K(u,s)$ is continuous over $\mathcal{U}$, which means it is also bounded since $\mathcal{U}$ is compact. Using Mercer's theorem it admits the spectral decomposition

$$K(u,s) = \sum_{k=1}^{\infty} \tau_k v_k(u) v_k(s). \tag{B.1}$$

We assume that eigenvalues decay as

$$\tau_k \asymp k^{-2h},$$

for some $h \geq 1$. Recall that, by Mercer's theorem, the convergence above occurs uniformly and absolutely in $u$ and $s$. We therefore have the following lemma, which will be used throughout.

**Lemma B.1.** *If $K(u,s)$ is a continuous, positive definite, and symmetric kernel then it admits the eigen-decomposition* (B.1)*, which satisfies*

$$\sup_{t,s} \tau_k |v_k(u) v_k(s)| \to 0 \quad \textit{as } k \to \infty.$$

The use of this Lemma B.1 is what allows us to relax the assumptions on the error process as compared to Cai and Yuan (2011), as it allows us to avoid certain Cauchy-Schwarz inequalities involving the errors (note it also fixes one misapplication of the Cauchy-Schwarz they had in their proofs). The functions $v_k(u)$ are normalized to have $L^2(\mathcal{U})$ norm one (from here on we notationally drop the domain $\mathcal{U}$), which also means they have $\mathbb{K}$ norm $\tau_k^{-1/2}$. Recall that the $\mathbb{K}$ inner product can be expressed as

$$\langle g, f \rangle_{\mathbb{K}} = \sum_{k=1}^{\infty} \frac{\langle f, v_k \rangle \langle g, v_k \rangle}{\tau_k},$$

where norms and inner products without subscripts will always denote the $L^2$ norm.

We now define the biased population parameter that will act as an intermediate value in our asymptotic derivation. First, define the population counterpart to $\mathbf{T}_{nm}$ from Section 4.2 as

$$[\mathbf{Tf}](u) := \mathrm{E}[K_{u_{11}}(u) \mathbf{\Sigma}_X \mathbf{f}(u_{11})] = \int K(u,s) \mathbf{\Sigma}_X \mathbf{f}(s) \, d\mu(s)$$

and $\mathbf{h} = \mathbf{T}(\boldsymbol{\beta}_0)$. We then define

$$\boldsymbol{\beta}_\lambda = (\mathbf{T} + \lambda \mathbf{I})^{-1} \mathbf{h} = (\mathbf{T} + \lambda I)^{-1} \mathbf{T} \boldsymbol{\beta}_0. \tag{B.2}$$

We now define a final intermediate value as

$$\tilde{\boldsymbol{\beta}}_\lambda = \boldsymbol{\beta}_\lambda + (\mathbf{T} + \lambda \mathbf{I})^{-1} (\mathbf{h}_{nm} - \mathbf{T}_{nm}(\boldsymbol{\beta}_\lambda) - \lambda \boldsymbol{\beta}_\lambda). \tag{B.3}$$

To establish our convergence rates we break up the problem into three pieces:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_0) + (\tilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda) + (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_\lambda).$$

In order to establish bounds for the third term above, it will be necessary to bound the second term in terms of the norm $\|f\|_\alpha = \langle K^{-\alpha/2} f, K^{-\alpha/2} f \rangle$. When $\alpha = 0$ this is the $L^2$ norm, when $\alpha = 1$ it is the $\mathbb{K}$ norm, but we allow intermediate values $\alpha \in [0, 1]$.

### *Step 1:* $\boldsymbol{\beta_\lambda - \beta_0}$

Using (B.2) we have

$$\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_0 = [(\mathbf{T} + \lambda \mathbf{I})^{-1} \mathbf{T} - \mathbf{I}]\boldsymbol{\beta}_0 = -\lambda (\mathbf{T} + \lambda \mathbf{I})^{-1} \boldsymbol{\beta}_0.$$

We want to compute the norm of this quantity in the product space $(L^2)^P$, which, equivalently, can be thought of as the tensor product space $\mathbb{R}^P \otimes L^2$. We can make this calculation cleaner by using an appropriate basis. In particular, recall that $v_k$ are the eigenfunctions of $K$, and we can add to them the eigenvectors of $\Sigma_X$, denoted as $\mathbf{u}_p$, we can then construct a basis for the space as

$$\{\mathbf{u}_p \otimes v_k : p = 1, \ldots, P \; k = 1, \ldots, \infty\}.$$

If we let $\eta_p$ denote eigenvalues of $\Sigma_X$, then the eigenvalues of $(\mathbf{T} + \lambda I)$ are $\eta_p \tau_k + \lambda$ and the eigenfunctions are $\mathbf{u}_p \otimes v_k$. Applying Parceval's identity yields

$$\begin{aligned}
\|\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_0\|^2 &= \sum_{p=1}^{P} \sum_{k=1}^{\infty} \langle \lambda (\mathbf{T} + \lambda \mathbf{I})^{-1} \boldsymbol{\beta}_0, \mathbf{u}_p \otimes v_k \rangle^2 \\
&= \lambda^2 \sum_p \sum_k \frac{1}{(\eta_p \tau_k + \lambda)^2} \langle \boldsymbol{\beta}_0, \mathbf{u}_p \otimes v_k \rangle^2 \\
&= \lambda^2 \sum_p \sum_k \frac{\tau_k}{(\eta_p \tau_k + \lambda)^2} \frac{\langle \boldsymbol{\beta}_0, \mathbf{u}_p \otimes v_k \rangle^2}{\tau_k} \\
&\leq \lambda^2 \|\boldsymbol{\beta}_0\|_{\mathbb{K}}^2 \sup_{p,k} \frac{\tau_k}{(\eta_p \tau_k + \lambda)^2} \leq \lambda^2 \nu \|\boldsymbol{\beta}_0\|_{\mathbb{K}}^2 \sup_{p,k} \frac{\eta_p \tau_k}{(\eta_p \tau_k + \lambda)^2}.
\end{aligned}$$

To bound the sup consider the function $f(x) = x^\gamma (x + \lambda)^{-2}$, over $x \geq 0$ and for some fixed $\gamma > 0$ (this level of generality will be useful later on). Notice that this function will attain its maximum at a finite value of $x$ if and only if $\gamma < 2$, for $\gamma \geq 2$ the maximum is attained at infinity. The derivative is given by

$$\gamma x^{\gamma - 1} (x + \lambda)^{-2} - 2x^\gamma (\lambda + x)^{-3}.$$

Setting equal to zero we have

$$\gamma(\lambda + x) - 2x = 0 \Longrightarrow x = \frac{\gamma}{2 - \gamma} \lambda.$$

So we have

$$\sup \frac{(\eta_p \tau_k)^\gamma}{(\eta_p \tau_k + \lambda)^2} \leq c_0 \lambda^{\gamma - 2}. \tag{B.4}$$

Note that throughout we take $c_0, c_1$, etc, to denote generic constants whose exact values may change depending on the context. Taking $\gamma = 1$ we conclude that

$$\|\boldsymbol{\beta}_\lambda - \boldsymbol{\beta}_0\|^2 \leq c_0 \lambda \nu \|\boldsymbol{\beta}_0\|^2. \tag{B.5}$$

***Step 2:*** $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}_{\boldsymbol{\lambda}}$

In this part we will bound the difference more generally using the $\alpha$ norm for $\alpha < 1 - 1/2h$. First, recall that, by definition of $\boldsymbol{\beta}_\lambda$ we have

$$\mathbf{T}\boldsymbol{\beta}_\lambda + \lambda \boldsymbol{\beta}_\lambda = \mathbf{h} \Longrightarrow \lambda \boldsymbol{\beta}_\lambda = \mathbf{h} - \mathbf{T}\boldsymbol{\beta}_\lambda = \mathbf{T}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_\lambda).$$

Plugging this into (B.3), the expression for $\tilde{\boldsymbol{\beta}}_\lambda$, we obtain

$$\tilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda = (\mathbf{T} + \lambda \mathbf{I})^{-1} \left[ \mathbf{h}_{nm} - \mathbf{T}_{nm}\boldsymbol{\beta}_\lambda - (\mathbf{T}\boldsymbol{\beta}_0 - \mathbf{T}\boldsymbol{\beta}_\lambda) \right].$$

This quantity has mean zero since, using (3.1) we have

$$\mathrm{E}[\mathbf{h}_{nm}](u) = \frac{1}{n} \sum_i \frac{1}{m_i} \sum_j \mathbf{X}_i \mathbf{X}_i^\top \mathrm{E}[\boldsymbol{\beta}(u_{ij}) K_{u_{ij}}(u)] = (\mathbf{T}\boldsymbol{\beta}_0)(u).$$

and using (3.2) we have

$$\mathrm{E}[\mathbf{T}_{nm}\boldsymbol{\beta}_\lambda](u) = (\mathbf{T}\boldsymbol{\beta}_\lambda)(u).$$

Using Parceval's identity we can express the expected difference in the $\alpha$ norm as

$$\mathrm{E}\,\|\tilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\|_\alpha^2 = \sum_p \sum_k \frac{1}{\tau_k^\alpha (\eta_p \tau_k + \lambda)^2} \mathrm{Var}(\langle \mathbf{h}_{nm} - \mathbf{T}_{nm}\boldsymbol{\beta}_\lambda, \mathbf{u}_p \otimes v_k \rangle).$$

Using the assumed independence across $i$ and the definitions (3.1) and (3.2) we have

$$\begin{aligned} &\mathrm{Var}(\langle \mathbf{h}_{nm} - \mathbf{T}_{nm}\boldsymbol{\beta}_\lambda, \mathbf{u}_p \otimes v_k \rangle) \\ &= \frac{1}{n^2} \sum_i \frac{1}{m_i^2} \mathrm{Var} \left( \sum_\ell (Y_{i\ell} - \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda(u_{i\ell})) \langle K_{u_{i\ell}}, v_k \rangle \mathbf{X}_i^\top \mathbf{u}_j \right). \end{aligned}$$

Using the reproducing property and that the $v_k$ are the eigenfunctions of $K$, we can express $\langle K_{u_{ij}}, v_k \rangle = \tau_k \langle K_{u_{ij}}, v_k \rangle_{\mathbb{K}} = \tau_k v_k(u_{ij})$. So the above is bounded by

$$\frac{\tau_k^2}{n^2} \sum_i \frac{(\mathbf{X}_i^\top \mathbf{u}_j)^2}{m_i^2} \mathrm{Var} \left( \sum_\ell (Y_{i\ell} - \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda(u_{i\ell})) v_k(u_{i\ell}) \right)$$

$$\leq \frac{\tau_k^2 P \zeta^2}{n^2} \sum_i \frac{1}{m_i^2} \operatorname{Var}\left( \sum_\ell (Y_{i\ell} - \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda(u_{i\ell})) v_k(u_{i\ell}) \right).$$

Conditioning on the sigma algebra generated by the locations, $\mathcal{F} = \sigma\{u_{ij}\}$, we get

$$\begin{aligned}
&\operatorname{Var}\left( \sum_j (Y_{ij} - X_i^\top \boldsymbol{\beta}_\lambda(u_{ij})) v_k(u_{ij}) \right) \\
&\quad = \operatorname{Var}\left( \mathrm{E}\left[ \sum_j (Y_{ij} - \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda(u_{ij})) v_k(u_{ij}) \middle| \mathcal{F} \right] \right) \\
&\qquad + \mathrm{E}\left[ \operatorname{Var}\left( \sum_j (Y_{ij} - \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda(u_{ij})) v_k(u_{ij}) \middle| \mathcal{F} \right) \right].
\end{aligned}$$

The first term is given by

$$\begin{aligned}
&\operatorname{Var}\left( \sum_j \mathbf{X}_i^\top (\boldsymbol{\beta}_0(u_{ij}) - \boldsymbol{\beta}_\lambda(u_{ij})) v_k(u_{ij}) \right) \\
&\quad = m_i \operatorname{Var}(\mathbf{X}_i^\top (\boldsymbol{\beta}_0(u_{11}) - \boldsymbol{\beta}_\lambda(u_{11})) v_k(u_{11})) \\
&\quad \leq m_i \mathrm{E}(\mathbf{X}_i^\top (\boldsymbol{\beta}_0(u_{11}) - \boldsymbol{\beta}_\lambda(u_{11})) v_k(u_{11}))^2 \\
&\quad = m_i \int [\mathbf{X}_i^\top (\boldsymbol{\beta}_0(u) - \boldsymbol{\beta}_\lambda(u))]^2 v_k(u)^2 \, d\mu(u) \\
&\quad \leq m_i |\mathbf{X}_i|^2 \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}_\lambda\|^2 \sup_u v_k(u)^2 \\
&\quad \leq c_0 P \zeta^2 m_i \tau_k^{-1} \lambda \|\boldsymbol{\beta}_0\|_{\mathbb{K}}^2.
\end{aligned}$$

Note the last line follows from Lemma B.1 and equation (B.5).

Turning to the second term, we have

$$\begin{aligned}
&\operatorname{Var}\left( \sum_j (Y_{ij} - \mathbf{X}_i^\top \boldsymbol{\beta}_\lambda(u_{ij})) v_k(u_{ij}) \middle| \mathcal{F} \right) \\
&\quad = \sum_{j\ell} \operatorname{Cov}(Y_{ij}, Y_{i\ell} | \mathcal{F}) v_k(u_{ij}) v_k(u_{i\ell}) \\
&\quad = \sum_{j\ell} (C(u_{ij}, u_{i\ell}) + \sigma^2 1_{j=\ell}) v_k(u_{ij}) v_k(u_{i\ell}).
\end{aligned}$$

When $j = \ell$ we use the assumed bounded variance and the orthonormality of the $v_k$ to obtain

$$\mathrm{E}[(C(u_{ij}, u_{ij}) + \sigma^2) v_k(u_{ij})^2] = \int C(u,u) v_k(u)^2 \, du + \sigma^2 \leq c_0.$$

When $j \neq \ell$ we use the definition of the covariance to obtain

$$\begin{aligned}\mathrm{E}[(C(u_{ij}, u_{i\ell})v_k(u_{ij})v_k(u_{i\ell})] &= \int\int v_k(u)C(u,s)v_k(s)\; dsdu \\ &= \langle v_k, Cv_k\rangle = \mathrm{E}\langle \varepsilon, v_k\rangle^2.\end{aligned}$$

Using generic $\{c_i\}$ for the constants and recalling that $m$ is the harmonic mean of the $m_i$ we get the bound

$$\begin{aligned}&\mathrm{E}\,\|\tilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\|_\alpha^2 \\ &\leq \sum_p \sum_k \frac{\tau_k^{2-\alpha}}{(\eta_p\tau_k+\lambda)^2}\frac{1}{n^2}\sum_{i=1}^n \frac{1}{m_i^2}\left[\frac{c_0 m_i\lambda}{\tau_k} + m_i c_1 + m_i^2\,\mathrm{E}\langle\varepsilon, v_k\rangle^2\right] \\ &= \sum_p \sum_k \frac{\tau_k^{2-\alpha}}{(\eta_p\tau_k+\lambda)^2}\frac{1}{n}\left[\frac{\lambda}{m\tau_k}c_0 + \frac{1}{m}c_1 + \mathrm{E}\langle\varepsilon, v_k\rangle^2\right]. \qquad \text{(B.6)}\end{aligned}$$

We bound each term in the summand separately. If $\tau_k \asymp k^{-2h}$ then so is $\eta_p\tau_k$, since $1 \leq p \leq P$. For an arbitrary $\gamma > 1/2h$ we have

$$\sum_{k=1}^\infty \frac{\tau_k^\gamma}{(\eta_p\tau_k+\lambda)^2} \asymp \int_0^\infty \frac{x^{-2h\gamma}}{(\lambda + x^{-2h})^2}\; dx = \int \frac{x^{2h(2-\gamma)}}{(\lambda x^{2h}+1)^2}\; dx.$$

Let $y = \lambda x^{2h}$ then $x = \lambda^{-1/2h}y^{1/2h}$ and $dx = \lambda^{-1/2h}(1/2h)y^{1/2h-1}dy$. Then the above becomes

$$\int \frac{\lambda^{-(2-\gamma)}y^{2-\gamma}}{(y+1)^2}\lambda^{-1/2h}(1/2h)y^{1/2h-1}dy = \frac{\lambda^{-(2-\gamma+1/2h)}}{2h}\int \frac{y^{1-\gamma+1/2h}}{(y+1)^2}dy.$$

Notice the integral is finite since $\gamma > 1/2h$. We therefore have that, for any $\gamma > 1/2h$ and $p = 1, \ldots, P$,

$$\sum_{k=1}^\infty \frac{\tau_k^\gamma}{(\eta_p\tau_k+\lambda)^2} \asymp \lambda^{-(2-\gamma+1/2h)}. \qquad \text{(B.7)}$$

Taking $\gamma = 1-\alpha$ and applying (B.7), which is greater than $1/2h$ as long as $\alpha < 1 - 1/2h$, the first term in (B.6) is given by

$$\sum_{p=1}^P \sum_{k=1}^\infty \frac{\tau_k^{1-\alpha}}{(\eta_p\tau_k+\lambda)^2}\frac{\lambda c_0}{nm} = O(\lambda^{-\alpha-1/2h}(nm)^{-1}).$$

Turning to the second term in (B.6), take $\gamma = 2-\alpha$ we have by the same arguments

$$\frac{c_2}{nm}\sum_p\sum_k \frac{\tau_k^{2-\alpha}}{(\eta_p\tau_k+\lambda)^2} \asymp (nm)^{-1}\lambda^{-\alpha-1/2h}.$$

Turning to the last term in (B.6) we can use that $\mathrm{E}\,\|\varepsilon\|^2 < \infty$ to obtain

$$\sum_p \sum_{k=1}^{\infty} \frac{\tau_k^{2-\alpha}}{(\eta_p \tau_k + \lambda)^2} \frac{1}{n} \mathrm{E}\langle \varepsilon, v_k \rangle^2 \leq \mathrm{E}\,\|\varepsilon\|^2 n^{-1} \nu^{2-\alpha} \max_k \frac{(\eta_p \tau_k)^{2-\alpha}}{(\tau_k + \lambda)^2}.$$

Applying (B.4) with $\gamma = 2 - \alpha$ we have that the above is equivalent to

$$\mathrm{E}\,\|\varepsilon\|^2 n^{-1} c_0 \lambda^{-\alpha},$$

We thus conclude that

$$\|\tilde{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_\lambda\|_\alpha^2 = O_P\left((nm)^{-1}\lambda^{-\alpha-1/2h} + n^{-1}\lambda^{-\alpha}\right).$$

There will be two values of $\alpha$ that are especially important. The first is when $\alpha = 0$, which we use to bound the $L^2$ norm, while the second is for an arbitrary $\alpha$ that satisfies $1/2h < \alpha < 1 - 1/2h$, as this will be used to bound the last term in the next subsection.

### *Step 3:* $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$

Recall that $\hat{\boldsymbol{\beta}} = (\mathbf{T}_{nm} + \lambda \mathbf{I})^{-1}\mathbf{h}_{nm}$ and $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_\lambda + (\mathbf{T} + \lambda \mathbf{I})^{-1}(\mathbf{h}_{nm} - \mathbf{T}_{nm}(\boldsymbol{\beta}_\lambda) - \lambda \boldsymbol{\beta}_\lambda)$. Note that this also implies that $\mathbf{h}_{nm} = (\mathbf{T}_{nm} + \lambda \mathbf{I})\hat{\boldsymbol{\beta}}$. So write

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda - (\mathbf{T} + \lambda \mathbf{I})^{-1}(\mathbf{h}_{nm} - \mathbf{T}_{nm}(\boldsymbol{\beta}_\lambda) - \lambda \boldsymbol{\beta}_\lambda) \\
&= (\mathbf{T} + \lambda \mathbf{I})^{-1}\left((\mathbf{T} + \lambda \mathbf{I})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda) - (\mathbf{h}_{nm} - (\lambda \mathbf{I} + \mathbf{T}_{nm})\boldsymbol{\beta}_\lambda)\right) \\
&= (\mathbf{T} + \lambda \mathbf{I})^{-1}\left((\mathbf{T} + \lambda \mathbf{I})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda) - (\mathbf{T}_{nm} + \lambda \mathbf{I})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda)\right).
\end{aligned}$$

Computing the $\alpha$ norm we can apply Parseval's and the definition of $\mathbf{T}_{nm}$ to obtain

$$\begin{aligned}
&\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_\alpha^2 \\
&= \sum_p \sum_k \frac{\tau_k^{-\alpha}}{(\eta_p \tau_k + \lambda)^2}\left[(\tau_k + \lambda)\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda, \mathbf{u}_p \otimes v_k \rangle - \langle (\mathbf{T}_{nm} + \lambda \mathbf{I})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda), \mathbf{u}_p \otimes v_k \rangle\right]^2 \\
&= \sum_p \sum_k \frac{\tau_k^{2-\alpha}}{(\eta_p \tau_k + \lambda)^2}\left[\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda, \mathbf{u}_p \otimes v_k \rangle - \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i} \mathbf{u}_p^\top(\hat{\boldsymbol{\beta}}(u_{ij}) - \boldsymbol{\beta}_\lambda(u_{ij}))v_k(u_{ij})\right]^2.
\end{aligned}$$

Notice that we can write $\hat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}_\lambda(u) = \sum_{\ell=1}^\infty \mathbf{h}_\ell v_\ell(u)$ where $h_{\ell p} = \langle \hat{g}_p - g_{\lambda,p}, v_\ell \rangle$. We can then write

$$\mathbf{u}_p^\top(\hat{\boldsymbol{\beta}}(u_{ij}) - \boldsymbol{\beta}_\lambda(u_{ij}))v_k(u_{ij}) = \sum_{\ell=1}^\infty \mathbf{u}_p^\top \mathbf{h}_\ell v_\ell(u_{ij})v_k(u_{ij}).$$

So the difference is given by

$$\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda, \mathbf{u}_p \otimes v_k \rangle - \frac{1}{n}\sum_{i=1}^n \frac{1}{m_i}\sum_{j=1}^{m_i} \mathbf{u}_p^\top(\hat{\boldsymbol{\beta}}(u_{ij}) - \boldsymbol{\beta}_\lambda(u_{ij}))v_k(u_{ij})$$

$$= \mathbf{u}_p^\top \mathbf{h}_k - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\sum_{\ell=1}^{\infty}\mathbf{u}_p^\top \mathbf{h}_\ell v_\ell(u_{ij})v_k(u_{ij})$$
$$= \sum_{\ell=1}^{\infty}\mathbf{u}_p^\top \mathbf{h}_\ell \left[\langle v_k, v_\ell\rangle - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} v_\ell(u_{ij})v_k(u_{ij})\right].$$

Let $\delta \in [0,1]$ be another constant similar, but potentially different from $\alpha$. We can then apply CS to bound the above by

$$\begin{aligned}
&|\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda, \mathbf{u}_p \otimes v_k\rangle| \\
&\quad\le \left(\sum_{\ell=1}^{\infty}\frac{(\mathbf{u}_p^\top \mathbf{h}_\ell)^2}{\tau_\ell^\delta}\right)\sum_{\ell=1}^{\infty}\tau_\ell^\delta\left[\langle v_k, v_\ell\rangle - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} v_\ell(u_{ij})v_k(u_{ij})\right]^2 \\
&\quad= \|\mathbf{u}_p^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda)\|_\delta^2\sum_{\ell=1}^{\infty}\tau_\ell^\delta\left[\langle v_k, v_\ell\rangle - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} v_\ell(u_{ij})v_k(u_{ij})\right]^2.
\end{aligned}$$

To get the asymptotic order of the summation term above, by Markov's inequality, it is enough to bound its expected value (since it is positive). Taking the expected value of the summation we get that

$$\begin{aligned}
&\sum_{\ell=1}^{\infty}\tau_\ell^\delta \,\mathrm{E}\left[\langle v_k, v_\ell\rangle - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i} v_\ell(u_{ij})v_k(u_{ij})\right]^2 \\
&= \sum_{\ell=1}^{\infty}\frac{\tau_\ell^\delta}{nm}\operatorname{Var}(v_\ell(u_{11})v_k(u_{11})) \\
&\le \sum_{\ell=1}^{\infty}\frac{\tau_\ell^\delta}{nm}\int v_\ell(u)^2 v_k(u)^2\, d\mu(u) \\
&\le \sum_{\ell=1}^{\infty}\frac{\tau_\ell^\delta}{nm}\sup_u v_k(u)^2\int v_\ell(u)^2\, du \\
&\le \sum_{\ell=1}^{\infty}\frac{c_0\tau_\ell^\delta}{nm\tau_k}.
\end{aligned}$$

Recall that $\tau_\ell \asymp \ell^{-2h}$, so the above sum is finite as long as $\delta > 1/2h$. Putting everything together and applying (B.4) we have the bound

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_\alpha^2 &\le O_P(1)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|_\delta^2\frac{c_0}{nm}\sum_k\frac{\tau_k^{1-\alpha}}{(\tau_k+\lambda)^2} \\
&\asymp O_P(1)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|_\delta^2(nm)^{-1}\lambda^{-\alpha-1/2h},
\end{aligned}$$

which holds for any $0 \le \alpha < 1 - 1/2h$ and any $\delta > 1/2h$.

Assume that $\lambda$ is such that $(nm)^{-1}\lambda^{-\alpha-1/2h} \to 0$, then it follows that $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_\alpha^2 = o_P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|_\delta^2)$. A triangle inequality gives

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|_\delta \geq \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|_\delta - \|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_\delta = (1 + o_P(1))\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|_\delta.$$

This implies that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|_\delta = O_P(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|_\delta).$$

Finally, take $\alpha = 0$ and $\delta > 1/2h$ then we have that

$$\begin{aligned}\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|^2 &= O_P(1)(nm)^{-1}\lambda^{-1/2h}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_\lambda\|_\delta^2 \\ &= O_P(1)(nm)^{-1}\lambda^{-1/2h}[(nm)^{-1}\lambda^{-\delta-1/2h} + n^{-1}\lambda^{-\delta}].\end{aligned}$$

If we assume that $\lambda$ is such that $(nm)^{-1}\lambda^{-\delta-1/2h} \to 0$ then the above simplifies to

$$o_P(1)\lambda^\delta[(nm)^{-1}\lambda^{-\delta-1/2h} + n^{-1}\lambda^{-\delta}] = o_P(1)[(nm)^{-1}\lambda^{-1/2h} + n^{-1}],$$

as desired.

Note that in the last paragraph, we made a more explicit assumption about how quickly $\lambda$ tends to zero. Note that the optimal rate is $\lambda = (nm)^{2h/(1+2h)}$. For this value of $\lambda$ we have that $(nm)^{-1}\lambda^{-\alpha-1/2h} \to 0$ for any value of $\alpha < 1$ since $1 + 1/2h = (2h+1)/2h$.

## Appendix C: Further on simulation

We discuss how we created the parabola domain and swiss roll domain for our simulation.

### *C.1. Parabola and swiss roll generation and geodesic distance*

*Parabola*

Domain is created using $t \to (x_1, x_2) = (t, t^2)$. For our simulation, I have used $t \in [-1, 1]$ so that $x_1 \in [-1, 1]$ and $x_2 \in [0, 1]$.

For finding the geodesic distance in parabola, the usual way is to use arc length.

$$\begin{aligned}L(t_1, t_2) &= \int_{t_1}^{t_2} \left\| \left(\frac{dx_1}{dt}\right)^2 + \left(\frac{dx_2}{dt}\right)^2 \right\| dt \\ &= \int_{t_1}^{t_2} \sqrt{1 + 4t^2} dt \\ &= \left[\frac{1}{4}\sinh^{-1}(2t) + \frac{1}{2}t\sqrt{1 + 4t^2}\right]_{t=t_1}^{t=t_2}.\end{aligned}$$

However, we did something special here. We have warped the time using the additional term $(1+4t^2)^{3/2}$ and changed the distance measure into the below.

$$\begin{aligned} L_2(t_1,t_2) &= \int_{t_1}^{t_2} (1+4t^2)^{3/2} \left\| \left(\frac{dx_1}{dt}\right)^2 + \left(\frac{dx_2}{dt}\right)^2 \right\| dt \\ &= \int_{t_1}^{t_2} (1+4t^2)^{3/2} \sqrt{1+4t^2} dt \\ &= \left[\frac{16}{5}t^5 + \frac{8}{3}t^3\right]_{t=t_1}^{t=t_2}. \end{aligned}$$

Therefore, if it is around $t=0$, then it would not be changed much, but when it's moving away from $t=0$, then the time warping will take effect and extends the distance much more than the regular arc length.

*Swiss roll*

The usual swiss roll can be created using $(t,s) \to (x_1,x_2,x_3) = (t\cdot\cos(t), s, t\cdot\sin(t))$. Since I have used $(t,s) \in [4,16]\times[4,16]$ to make a nice shape of swiss roll, I have scaled it using

$$(t,s) \to (x_1,x_2,x_3) = \frac{1}{15}(t\cdot\cos(t), s, t\cdot\sin(t))$$

to make $x_1 \in (-1.1, 1)$, $x_2 \in (0.25, 1.1)$, and $x_3 \in (-1,1)$.

Now let's find the geodesic distance in swiss roll. We know that swiss roll is basically a plane rolled in, and we will use this information. First we find the distance over $(x_1, x_3)$.

$$\begin{aligned} L_t(t_1,t_2) &= \int_{t_1}^{t_2} \left\| \left(\frac{dx_1}{dt}\right)^2 + \left(\frac{dx_3}{dt}\right)^2 \right\| dt \\ &= \int_{t_1}^{t_2} \left(\frac{1}{15}\right) \sqrt{(\cos(t) - t\sin(t))^2 + (\sin(t) + x\cos(t))^2} dt \\ &= \frac{1}{15}\int_{t_1}^{t_2} \sqrt{t^2+1} dt \\ &= \frac{1}{15}\left[\frac{1}{2}\sinh^{-1}(t) + \frac{1}{2}t\sqrt{1+t^2}\right]_{t=t_1}^{t=t_2}. \end{aligned}$$

Now we have $L_t$, we will use this to find the geodesic distance on swiss roll.

$$d_{\mathcal{S}}\left((t_1,s_1),(t_2,s_2)\right) = \sqrt{L_t(t_1,t_2)^2 + (s_2-s_1)^2/15^2}.$$

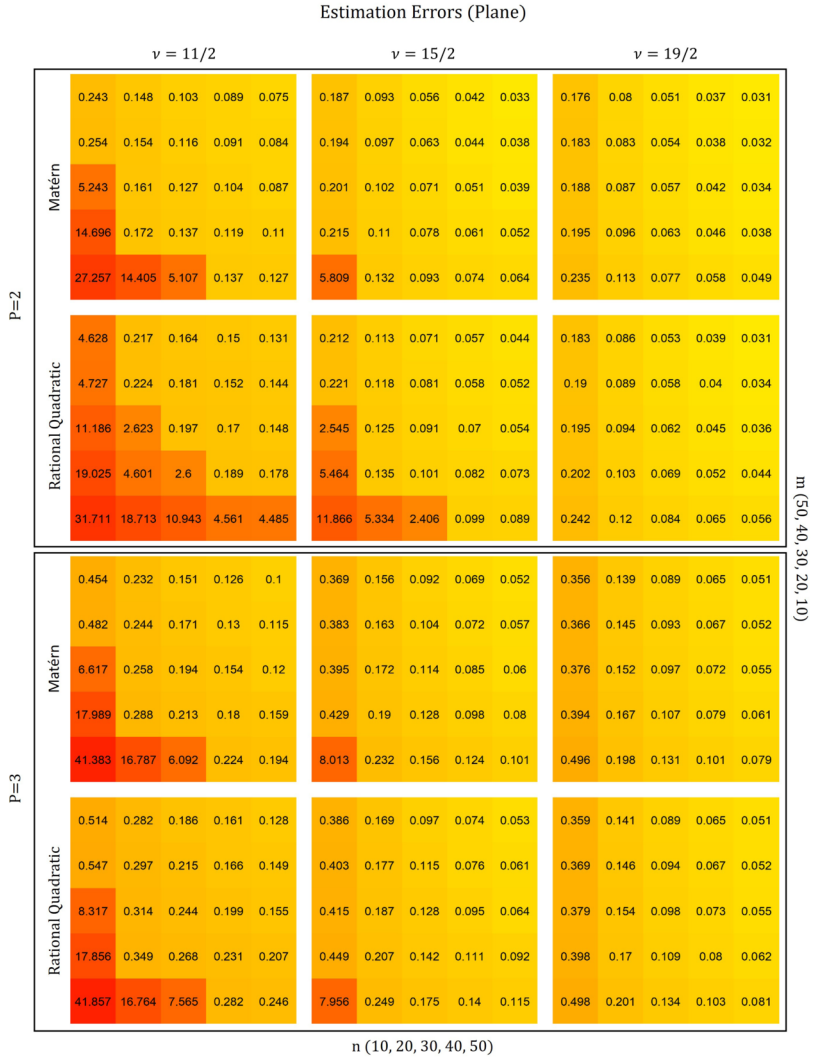


FIG 10. *The effects of $n$ and $m$ on the mean squared estimation errors for plane domain with $P = 2$ and $P = 3$ cases.*

### *C.2. More simulation results*

The estimation errors for plane $\mathcal{U}$ is presented in Figure 10, and the estimation errors for parabola $\mathcal{U}$ is presented in Figure 11. These show similar trends as discussed in the main manuscript. The estimation errors drop as $n$ increases and $m$ increases, but at some point of $m$, they do not change as much because $n$ dominates the rate. For example, in the setting b1C (plane domain, $P = 2$, and $\nu = 11/2$), the estimation errors drop from 5.107 to 0.137 as $m$ in increases

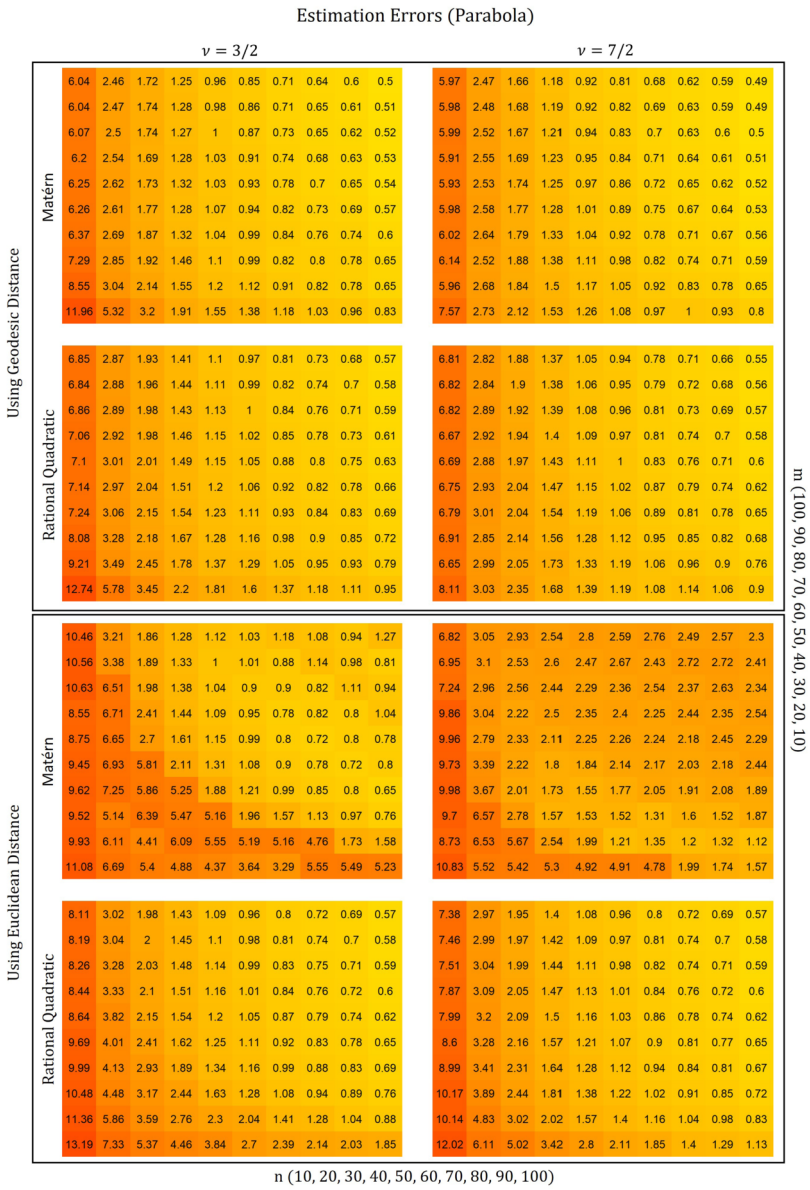


FIG 11. *The effects of $n$ and $m$ on the mean squared estimation errors for parabola domain with using geodesic distance and with using Euclidean distance.*

from 10 to 20 and $n = 30$, but as $m$ increases from 20 to 30, the estimation error change from 0.137 to 0.127.

It is clearly shown that the estimation really fails using Euclidean distance for parabola $\mathcal{U}$. When we used geodesic distance, the estimation errors using

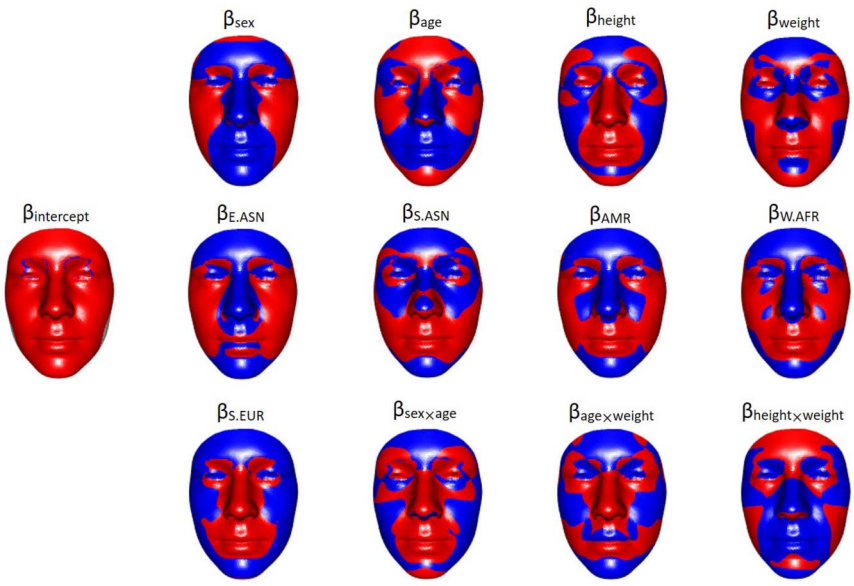


FIG 12. *The estimated beta plots for 3D facial data. The red means outward effect and the blue means inward effect.*

Matérn kernel and the estimation errors using rational quadratic kernel have similar values and show similar trends, but when we used Euclidean distance, the estimation error using Matérn kernel especially show failure.

## Appendix D: Estimation results of 3D facial data

In Figure 12, we have plotted the 13 estimated beta plots for ADAPT 3D facial data. The red shows outward effect and the blue shows inward effect. The effect of sex, the effect of age, the effect of Eastern Asian, and the effect of Southern African are discussed in the main manuscript. The parameter for height shows that as a person gets taller, the cheek areas would decrease whereas the chin will be outer, showing overall longer face. For the parameter for weight, the outer effect in the middle of nose with inward effect on the top and on the bottom of the nose show rounder nose as weight increases, and the inward effect below the lip shows more labiomental groove, and the outward effect on the chin can mean more fat on the jaw.

One thing to note is that our model is a linear model, and the effects captured here are linear. However, some of the effects of the predictors, for example age, can be nonlinearlly affecting the shape of face, and our current model does not capture that. We plan on capturing such nonlinear effects in the future study.

## References


[1] Aronszajn, N. and Smith, K. T. (1961), Theory of Bessel potentials. i, *in* 'Annales de l'institut Fourier', Vol. 11, pp. 385–475. MR0143935

[2] Barber, R. F., Reimherr, M., Schill, T. et al. (2017), 'The function-on-scalar lasso with applications to longitudinal gwas', *Electronic Journal of Statistics* **11**(1), 1351–1389. MR3635916

[3] Berlinet, A. and Thomas-Agnan, C. (2011), *Reproducing kernel Hilbert spaces in Probability and Statistics*, Springer Science & Business Media. MR2239907

[4] Cai, T. T. and Yuan, M. (2011), 'Optimal estimation of the mean function based on discretely sampled functional data: Phase transition', *The Annals of Statistics* **39**(5), 2330–2355. MR2906870

[5] Cai, T. T. and Yuan, M. (2012), 'Minimax and adaptive prediction for functional linear regression', *Journal of the American Statistical Association* **107**(499), 1201–1216. MR3010906

[6] Canzani, Y. (2013), 'Analysis on manifolds via the Laplacian', *Lecture Notes available at: http://www.math.harvard.edu/canzani/docs/Laplacian.pdf. Google Scholar*.

[7] Cho, Y.-K. (2017), 'Compactly supported reproducing kernels for $l^2$-based Sobolev spaces and Hankel-Schoenberg transforms', *arXiv preprint arXiv:1702.05896*.

[8] Choe, A. S., Nebel, M. B., Barber, A. D., Cohen, J. R., Xu, Y., Pekar, J. J., Caffo, B. and Lindquist, M. A. (2017), 'Comparing test-retest reliability of dynamic functional connectivity methods', *Neuroimage* **158**, 155–175.

[9] Claes, P., Hill, H. and Shriver, M. D. (2014a), 'Toward dna-based facial composites: preliminary results and validation', *Forensic Sci Int Genet* **13**, 208–16.

[10] Claes, P., Liberton, D. K., Daniels, K., Rosana, K. M., Quillen, E. E., Pearson, L. N., McEvoy, B., Bauchet, M., Zaidi, A. A., Yao, W., Tang, H., Barsh, G. S., Absher, D. M.,... and Shriver, M. D. (2014b), 'Modeling 3D facial shape from DNA', *PLoS Genet* **10(3)**.

[11] Craioveanu, M.-E., Puta, M. and Rassias, T. (2013), *Old and New Aspects in Spectral Geometry*, Vol. 534, Springer Science & Business Media. MR1880186

[12] Dai, X., Müller, H.-G. et al. (2018), 'Principal component analysis for functional data on Riemannian manifolds and spheres', *The Annals of Statistics* **46**(6B), 3334–3361. MR3852654

[13] 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', *Journal of Multivariate Analysis* **12**(1), 136–154. MR0650934

[14] Duchi, J. (2016), 'Lecture notes for statistics 311/electrical engineering 377', *URL: https://stanford.edu/class/stats311/Lectures/full_notes.pdf. Last visited on* **2**, 23.

[15] Edmunds, D. E. and Triebel, H. (1996), *Function Spaces, Entropy Numbers,*

*Differential Operators*, Cambridge University Press. MR1410258
[16] Ettinger, B., Perotto, S. and Sangalli, L. M. (2016), 'Spatial regression models over two-dimensional manifolds', *Biometrika* **103**(1), 71–88. MR3465822
[17] Fan, Z. and Reimherr, M. (2017), 'High-dimensional adaptive function-on-scalar regression', *Econometrics and Statistics* **1**, 167–183. MR3669995
[18] Hall, P., Horowitz, J. L. et al. (2007), 'Methodology and convergence rates for functional linear regression', *The Annals of Statistics* **35**(1), 70–91. MR2332269
[19] Hebey, E. (2000), *Nonlinear analysis on manifolds: Sobolev spaces and inequalities*, Vol. 5, American Mathematical Soc. MR1688256
[20] Jayasumana, S., Hartley, R., Salzmann, M., Li, H. and Harandi, M. (2013), Kernel methods on the riemannian manifold of symmetric positive definite matrices, *in* 'proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 73–80.
[21] Jirak, M. (2016), 'Optimal eigen expansions and uniform bounds', *Probability Theory and Related Fields* **166**(3-4), 753–799. MR3568039
[22] Kang, H. B., Reimherr, M., Shriver, M. and Claes, P. (2017), 'Manifold data analysis with applications to high-frequency 3D imaging', *arXiv preprint arXiv:1710.01619*.
[23] Lee, W., Miranda, M. F., Rausch, P., Baladandayuthapani, V., Fazio, M., Downs, J. C. and Morris, J. S. (2018), 'Bayesian semiparametric functional mixed models for serially correlated functional data, with application to glaucoma data', *Journal of the American Statistical Association*. MR3963158
[24] Li, Y., Hsing, T. et al. (2010), 'Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data', *The Annals of Statistics* **38**(6), 3321–3351. MR2766854
[25] Lila, E., Aston, J. A., Sangalli, L. M. et al. (2016), 'Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging', *The Annals of Applied Statistics* **10**(4), 1854–1879. MR3592040
[26] Lin, Z. and Yao, F. (2018), 'Intrinsic Riemannian functional data analysis', *arXiv preprint arXiv:1812.01831*. MR4025751
[27] Pazouki, M. and Schaback, R. (2011), 'Bases for kernel-based spaces', *Journal of Computational and Applied Mathematics* **236**(4), 575–588. MR2843040
[28] Petrovich, J. and Reimherr, M. (2017), 'Asymptotic properties of principal component projections with repeated eigenvalues', *Statistics & Probability Letters* **130**, 42–48. MR3692217
[29] Reimherr, M., Sriperumbudur, B. and Taoufik, B. (2017), 'Optimal prediction for additive function-on-function regression', *arXiv preprint arXiv:1708.03372*. MR3893421
[30] Sun, X., Du, P., Wang, X. and Ma, P. (2018), 'Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework', *Journal of the American Statistical Association* **113**(524), 1601–1611. MR3902232
[31] Varshamov, R. (1957), 'Estimate of the number of signals in error correcting

codes', *Docklady Akad. Nauk, SSSR* **117**, 739–741. MR0095090

[32] Wahba, G. (1990), *Spline Models for Observational Data*, Vol. 59, Siam. MR1045442

[33] Wang, X. and Ruppert, D. (2015), 'Optimal prediction in an additive functional model', *Statistica Sinica* pp. 567–589. MR3379089

[34] Zhang, X. and Wang, J.-L. (2018), 'Optimal weighting schemes for longitudinal and functional data', *Statistics & Probability Letters* **138**, 165–170. MR3788733

[35] Zhang, X., Wang, J.-L. et al. (2016), 'From sparse to dense functional data and beyond', *The Annals of Statistics* **44**(5), 2281–2321. MR3546451