Statistica Sinica Preprint No: SS-2024-0253	
Title	Predictive Distributions and the Transition from Sparse to Dense Functional Data
Manuscript ID	SS-2024-0253
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0253
Complete List of Authors	Álvaro Gajardo, Xiongtao Dai and Hans-Georg Müller
Corresponding Authors	Xiongtao Dai
E-mails	xiongtao.dai@hotmail.com
Notice: Accepted author version.	

Predictive Distributions and the Transition from Sparse to Dense Functional Data

Álvaro Gajardo[1], Xiongtao Dai[2], and Hans-Georg Müller[1]

[1]*Department of Statistics, University of California, Davis, USA*

[2]*Division of Biostatistics, University of California, Berkeley, USA*

Abstract: Gaussian distributed sparsely sampled longitudinal data can be represented as Gaussian distributions of their functional principal component scores, conditional on the available data. Since these conditional distribu-tions reflect the entire information available about these scores and therefore about the unknown trajectories that constitute the realizations of the stochastic process that generates the functional data, they are referred to as predictive distributions. This motivates a deeper investigation of the convergence of the predicted functional principal component scores given noisy longitudinal observations towards the true but unobservable scores as the designs transition from sparse (longitudinal) to dense (functional) and of the shrinkage of the predictive distributions towards a point mass located at the true score as the number of observations per subject increases. Our study is motivated by the theoretical and practically relevant challenge that point predictions in the sparse sampling regime are not consistent for the true functional principal component scores. Our proposal is to change the perspective towards a focus on predictive distributions, which can be consistently estimated. The emphasis is thus shifted to uncertainty quantification. This approach is also demonstrated for the case of sparsely sampled longitudinal predictors in functional linear models where again one does not have consistent point predictors. Theoretical justification is provided through the asymptotic rates of convergence for the 2-Wasserstein metric between true and estimated predictive distributions. The application of the predictive distribution approach for functional principal component analysis is illustrated for longitudinal data from the Baltimore Longitudinal Study of Aging.

Key words and phrases: Functional Data Analysis, Functional Principal Components, Functional Regression, Longitudinal Data, Sparse Design, Sparse-to-Dense, Uncertainty Quantification, Wasserstein Metric.

This research was done while Álvaro Gajardo was a PhD student at the University of California, Davis.

Authors' ORCID
Álvaro Gajardo 0000-0002-5507-7588
Xiongtao Dai 0000-0002-6996-5930
Hans-Georg Müller 0000-0002-6396-2552

1. Introduction

1.1 General perspective and background

Functional Data Analysis has found a wide range of applications (Ramsay and Silverman, 2005; Horvath and Kokoszka, 2012; Wang *et al.*, 2016). These include longitudinal studies, where functional principal component analysis (Kleffe, 1973; Castro *et al.*, 1986), a core technique of Functional Data Analysis, was shown to play a central role, due to its interpretability and ease of implementation. A key feature of many longitudinal studies is the sparsity of the available observations per subject, which are inherently correlated and are often available at only a few irregular times and usually contaminated with measurement error.

When subjects are recorded densely over time, one can consistently recover underlying random trajectories from the Karhunen–Loève representation. Starting from the auto-covariance function of the process X given by

$$\Gamma(s,t) = \text{cov}(X(s),X(t)) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t), \quad s,t \in \mathcal{T}, \tag{1.1}$$

where $\lambda_1 > \lambda_2 > \cdots \geq 0$ are the ordered eigenvalues, satisfying $\sum_{k=1}^{\infty} \lambda_k < \infty$, and ϕ_k, $k \geq 1$, are the orthonormal eigenfunctions associated with the Hilbert–Schmidt operator $\Xi(g) = \int_{\mathcal{T}} \Gamma(\cdot,t)g(t)dt$, $g \in L^2(\mathcal{T})$. Define eigengaps $\delta_k = \min(\lambda_{k-1} - \lambda_k, \lambda_k - \lambda_{k+1})$, $k = 1, 2, \ldots$, and denote by $\mu(t) = E(X_i(t))$ the mean function, by $X_i^c(t) = X_i(t) - \mu(t)$ the centered process, and by $\xi_{ik} = \int_{\mathcal{T}} X_i^c(t)\phi_k(t)dt$ the kth functional principal component, $k = 1, 2, \ldots$, which satisfies $E(\xi_{ik}) = 0$, $E(\xi_{ik}^2) = \lambda_k$ and $E(\xi_{ik}\xi_{il}) = 0$ for $k, l = 1, 2, \ldots$, $l \neq k$. Trajectories can then be represented through the Karhunen–Loève decomposition, also referred to as functional principal component analysis (FPCA),

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t), \tag{1.2}$$

where in practice it is often useful to consider a truncated expansion using the first $K > 0$ components that explain most of the variation, for example through the fraction of variance explained or FVE criterion (Yao *et al.*, 2005a).

A common approach is to employ Riemann sums to recover the integrals that represent the projections of the trajectories on the eigenfunctions of the auto-covariance operator of the underlying stochastic process. These integrals correspond to the functional principal components and their approximation by Riemann sums is known to improve as the number of observations per subject increases (Müller, 2005). However, when functional data are sparsely observed, which means that only a finite number of observations are available for each subject, this approximation is not feasible. To address this challenge, Yao *et al.* (2005a) introduced the Principal Analysis through Conditional Expectation (PACE) approach, which aims to recover the underlying trajectories by targeting the best predictions conditional on the observations under Gaussianity assumptions and otherwise the best linear predictor. A related approach has been Gaussian Processes (Shi and Choi, 2011; Wang and Shi, 2014). Best predictions of the functional principal components can be consistently estimated based on consistent nonparametric estimates of mean and covariance functions that are obtained by pooling all observations across subjects, borrowing strength from the entire sample. While these best predictions are unbiased, they do not lead to consistent trajectory recovery (Yao *et al.*, 2005a).

A second scenario where consistent predictions are unavailable in the sparse case when one has only a finite number of observations per subject is the Functional Linear Regression Model for the relationship between a scalar or functional response Y and functional predictors $X(t)$, $t \in \mathcal{T}$, a compact interval (Ramsay and Silverman, 2005; Hall and Horowitz, 2007; Shi and Choi, 2011; Kneip *et al.*, 2016; Chiou *et al.*, 2016),

$$E[Y|X] = \mu_Y + \int_{\mathcal{T}} \beta(t) X^c(t) dt. \tag{1.3}$$

Here $\mu_Y = E(Y)$, $X^c(t) = X(t) - E(X(t))$ and the slope function β lies in $L^2(\mathcal{T})$. The unavailability of consistent predictions in the functional linear model is a consequence of the fact that the integral appearing in (1.3) cannot be consistently approximated in the sparse sampling case, even in the case where the slope function β is known. In contrast to the prediction task, a consistent estimate of the slope function β in model (1.3) can be obtained through consistent cross-covariance estimation by pooling the sparse data (Yao *et al.*, 2005b).

The behavior of estimating mean functions and covariance functions, as well as the associated eigenvalues and eigenfunctions as per (1.1) in dense and sparse cases has been the subject of numerous studies (Cai and Hall, 2006; Hall and Horowitz, 2007; Chiou *et al.*, 2016; Kneip *et al.*, 2016; Hall and Hosseini-Nasab, 2006; Müller and Yao, 2010). Specifically, the effect of the transition from sparse to dense designs on the convergence rates for the estimation of mean functions, auto-covariance functions and cross-covariance functions and related phase transitions have been studied in detail (Li and Hsing, 2010; Zhang and Wang, 2016, 2018). However, there are only very few studies about the sparse to dense behavior of estimates of the principal components ξ in FPCA (1.2) (Müller, 2005; Dai *et al.*, 2018) and we are not aware of any study about the behavior of predictions in the FLM (1.3).

1.2 Innovation and outline of proposed approach

In this paper we address the challenge of obtaining consistent predictions for trajectories or responses when one has sparse data in the context of functional principal component analysis (FPCA)

(1.1) or the functional linear model (FLM) (1.3). Specifically, this work includes (1) The study of the behavior of functional principal components predicted from data when the data sampling transitions from the sparse to the dense case, which complements previous studies on the behavior of mean and covariance function estimates under this transition; (2) the idea to replace the inherently inconsistent point estimates of functional principal components and of responses in functional linear models under sparse sampling by consistent estimates of predictive distributions that correspond to the conditional distributions of outcomes of interest (functional trajectories or predicted responses); these distributions indicate where the quantities to be predicted are likely situated based on the available data without providing a precise location; of interest is the shrinkage of these predictive distributions in the transition from sparse to dense designs; (3) consistent estimates for predictive distributions and their convergence to the true predictive distributions.

Our proposal is to rephrase the prediction problem for trajectories in the FPCA case and of scalar outcomes in the FLM case by shifting the target from point prediction, i.e., the problem of predicting conditional expectations for which consistency is unachievable, to the problem of estimating a predictive distribution, i.e., a conditional distribution rather than a conditional expectation. This new perspective leads to a target for which consistent estimation is indeed feasible. To study the behavior of predictive distributions under Gaussian assumptions, it proves useful to consider a map from sparse and irregularly sampled data to a multivariate Gaussian predictive distribution for a vector of truncated functional principal components and to investigate its behavior as the number of observations per subject increases.

One of our goals is to quantify the accompanying shrinkage of the conditional predictive distributions given the data and their convergence towards a point mass located at the true but unobserved functional component scores. To predict the expected response $E[Y|X]$ in model (1.3) in the sparse case, a feasible approach is to construct predictive distributions for the expected response given the information available for a subject. These predictive distributions can be consistently estimated in both the Wasserstein and Kolmogorov metric (Villani, 2003) and we adopt a Wasserstein discrepancy measure to assess the predictability of the response by the predictive distribution. This measure is interpretable, can be consistently recovered under mild assumptions and is supported in simulations for various sparse designs and noise levels.

The paper is structured as follows: Preliminary results are in Section 2, where the convergence of the best predicted FPCs towards the true unobserved FPCs is established when transitioning from sparse to dense data. Crucially, this study does not require distributional assumptions. The concept of representing sparse functional/longitudinal data by predictive distributions for the FPCs in the case of Gaussian processes and our main results are the theme of Section 3, followed by an analysis of the shrinkage of the predictive distributions towards a point mass located at the true scores. Extensions to the shrinkage of the entire functional predictive distribution in the 2-Wasserstein metric are also presented in Section 3. This is followed by a study of the prediction of scalar responses Y in a Functional Linear Model (1.3) when predictors are sparsely observed in Section 4, extending the concept of predictive distributions for the predictable part of the response and assessing the predictability of Y by the predictive distribution through a Wasserstein discrepancy measure. Asymptotic results for the consistent estimation of both the predictive distributions and Wasserstein discrepancy in the sparse case are presented and a study of the behavior of the predictive distribution in the transition from sparse to dense sampling is also included. This is followed by simulation results in Section 5 to demonstrate the finite sample performance of the proposed methods. Finally, data illustrations for the proposed predictive distributions are presented in Section 6. The paper concludes with a discussion of the new predictive perspective in Section

6. Proofs and auxiliary results can be found in the Supplement.

2. Convergence of Predicted Functional Principal Components When Transitioning from Sparse to Dense Sampling

Assume that for each individual $i = 1, \dots, n$, there is an underlying unobserved function $X_i(t)$, where the functions X_i are i.i.d. realizations of a L^2-stochastic process $X(t)$, $t \in \mathcal{T}$, and $\mathcal{T}$ is a closed and bounded interval on the real line. Without loss of generality let $\mathcal{T} = [0, 1]$. Sparsely sampled and error-contaminated observations $\tilde{X}_{ij} = X_i(T_{ij}) + \epsilon_{ij}$, $j = 1, \dots, n_i$, $n_i \leq N_0$ for a finite N_0 are obtained at random times $T_{ij} \in \mathcal{T}$ that are distributed according to a continuous smooth distribution F_T, and write $\mathbf{T}_i = (T_{i1}, \dots, T_{in_i})^T$ to denote the vector of sampling time points for the ith subject. The following condition is required:

(X1) $\{T_{ij} : i = 1, \dots, n, \, j = 1, \dots, n_i\}$ are i.i.d. copies of a random variable T defined on $\mathcal{T}$, and n_i are non-random. The density $f(\cdot)$ of T is bounded below, $\min_{t \in \mathcal{T}} f(t) \geq m_f > 0$.

Assumption (X1) is standard (Zhang and Wang, 2016; Dai *et al.*, 2018) to ensure there are no systematic sampling gaps. The measurement errors ϵ_{ij} are assumed to be i.i.d. with mean 0 and variance σ^2, and independent of the underlying process $X_i(\cdot)$. We note that the results presented in this section do not require Gaussianity or other specific distributional assumptions. Throughout, our analysis is conditional on the random number of observations per subject n_i (Zhang and Wang, 2016).

Denoting by $\mathbf{T}_i = (T_{i1}, \dots, T_{in_i})^T$ the sampling time points for the ith subject and writing $\mathbf{X}_i = (\tilde{X}_{i1}, \dots, \tilde{X}_{in_i})^T$ and conditional on $\mathbf{T}_i$, it follows from (1.1) and (1.2) that $\text{cov}(\tilde{X}_{ij}, \xi_{ik} | \mathbf{T}_i) =$

$\lambda_k\phi_k(T_{ij})$, $j = 1, \ldots, n_i$, $k = 1, \ldots, K$. Define

$$\boldsymbol{\Phi}_{iK} = \begin{pmatrix} \phi_1(T_{i1}) & \ldots & \phi_K(T_{i1}) \\ \vdots & \vdots & \vdots \\ \phi_1(T_{in_i}) & \ldots & \phi_K(T_{in_i}) \end{pmatrix},$$

$\boldsymbol{\mu}_i = \mathrm{E}(\mathbf{X}_i|\mathbf{T}_i) = (\mu(T_{i1}), \ldots, \mu(T_{in_i}))^T$ and the $n_i \times n_i$ conditional covariance matrix $\boldsymbol{\Sigma}_i = \mathrm{cov}(\mathbf{X}_i|\mathbf{T}_i)$, for which the (j, l) entry is given by $\sigma^2\delta_{jl} + \Gamma(T_{ij}, T_{il})$, where $\delta_{jl} = 1$ if $j = l$ and 0 otherwise. To predict the functional principal components $\boldsymbol{\xi}_{iK} = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{iK})^T$, we utilize best linear unbiased predictors (Rice and Wu, 2001) of $\boldsymbol{\xi}_{iK}$ given $\mathbf{X}_i$ and $\mathbf{T}_i$, which are given by

$$\tilde{\boldsymbol{\xi}}_{iK} = \boldsymbol{\Lambda}_K \boldsymbol{\Phi}_{iK}^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_i) \text{ with } \boldsymbol{\Lambda}_K = \mathrm{diag}(\lambda_1, \ldots, \lambda_K). \tag{2.4}$$

As the number of observations for an individual increases as the functional sampling gets denser, the predicted functional principal components $\tilde{\boldsymbol{\xi}}_{iK}$ converge to their targets $\boldsymbol{\xi}_{iK}$ under the following assumptions.

(X2) The process $X(t)$ is continuously differentiable a.s. for $t \in \mathcal{T}$.

(X3) $\partial\Gamma(s, t)/\partial s$ exists and is continuous, for $s, t \in \mathcal{T}$.

Assumptions (X2)–(X3) are requirements for the smoothness of the original process and the covariance function, respectively. The following result does not require Gaussian assumptions.

Proposition 1. *Suppose that (X1)–(X3) hold and the number of observations n_i for the ith subject satisfies $n_i = m \to \infty$, $i = 1, \ldots, n$. Then, for any fixed $K \geq 1$, $k = 1, \ldots, K$, and $i = 1, \ldots, n$, as $m \to \infty$,*

$$|\tilde{\xi}_{ik} - \xi_{ik}| = O_p(m^{-1/2}). \tag{2.5}$$

Note that this result is for increasingly dense sampling across all subjects and indicates how this leads to better approximation of ξ_{ik}. The rate of convergence is the same as derived previously in Dai *et al.* (2018) for the functional principal components of the derivative process $X'(t)$ under Gaussian assumptions. This previous analysis utilized convergence results for nonparametric posterior distributions (Shen, 2002) that are tied to the Gaussian assumption, whereas this paper presents a novel direct approach that does not require distributional assumptions on X.

Next, we study scenarios where the unknown population quantities are estimated from the available data, and the subjects are assumed to be observed either on dense designs, with $n_i = m \to \infty$, or on sparse designs, with $2 \leq n_i \leq N_0 < \infty$ for a fixed number $N_0 < \infty$. Consider sequences

$$\begin{aligned} a_{n1} &= h_\mu^2 + \left\{ \frac{\log(n)}{nh_\mu} \right\}^{1/2}, \quad b_{n1} = h_G^2 + \left\{ \frac{\log(n)}{nh_G^2} \right\}^{1/2}, \\ a_{n2} &= h_\mu^2 + \left\{ \left(1 + \frac{1}{mh_\mu}\right) \frac{\log(n)}{n} \right\}^{1/2}, \quad b_{n2} = h_G^2 + \left(1 + \frac{1}{mh_G}\right) \left\{ \frac{\log(n)}{n} \right\}^{1/2} \end{aligned} \tag{2.6}$$

with bandwidths h_μ and h_G. For sparse designs, define sequences $a_n = a_{n1}$ and $b_n = b_{n1}$, while for dense designs these sequences will be defined as $a_n = a_{n2}$ and $b_n = b_{n2}$. Note that for dense designs the rates a_n and b_n also depend on m.

The estimation of mean function μ and covariance surface Γ utilizes local linear smoothers, in analogy to Zhang and Wang (2016), with further details in the Supplement Section S2.1. For the covariance smoothing step $n_i \geq 2$ is assumed throughout. The estimation of remaining population quantities such as σ^2 and eigenpairs (λ_k, ϕ_k), $k \geq 1$, is carried out analogously as in equations (2) and (3) in Yao *et al.* (2005a). Denote by $\hat{\Xi}$ the estimated counterpart of the Hilbert–Schmidt integral operator Ξ with eigenpairs $(\hat{\lambda}_k, \hat{\phi}_k)$ such that $\langle \hat{\phi}_k, \phi_k \rangle_{L^2} \geq 0$, where $\langle \cdot, \cdot \rangle_{L^2}$ denotes the L^2 inner product and $k \geq 1$.

Consider the estimated functional principal components in (1.2) for a new independent subject

i^* that is not part of the training data sample ($i = 1, \ldots, n$) and for which measurements are available over a dense but possibly irregular grid. Then as the design gets denser, these estimates converge to the true functional principal components, irrespective of whether the subjects in the training set are observed under sparse or dense designs. Specifically, for a realization X^* of the process X that is independent of $X_1, \ldots, X_n$, assume one has measurements of the process X^* at times T_j^* ($j = 1, \ldots, m^*$) with added noise $\mathbf{X}^* = (X^*(T_1^*) + \epsilon_1^*, \ldots, X^*(T_{m^*}^*) + \epsilon_{m^*}^*)$. Here $m^* \to \infty$ and the errors ϵ_j^* have mean zero and variance σ^2 and are independent of all other random quantities. Consider estimates $\hat{\xi}_k^* = \hat{\lambda}_k \hat{\phi}_k(\mathbf{T}^*)^T \hat{\mathbf{\Sigma}}^{*-1}(\mathbf{X}^* - \hat{\boldsymbol{\mu}}^*)$, where $\hat{\boldsymbol{\mu}}^* = \hat{\mu}(\mathbf{T}^*) := (\hat{\mu}(T_1^*), \ldots, \hat{\mu}(T_{m^*}^*))^T$, $\hat{\phi}_k(\mathbf{T}^*) = (\hat{\phi}_k(T_1^*), \ldots, \hat{\phi}_k(T_{m^*}^*))^T$, $\mathbf{T}^* = (T_1^*, \ldots, T_{m^*}^*)^T$, and $\hat{\mathbf{\Sigma}}^{*-1}$ is analogous to $\mathbf{\Sigma}_i^{-1}$ but replacing the T_{ij} with T_j^* and the population quantities by their estimated counterparts. A requirement is that

(B1) The eigenvalues $\lambda_1 > \lambda_2 > \cdots > 0$ are all distinct.

Theorem 1. *Suppose that assumptions (X2), (B1) and (A1)–(A8) in the Supplement Section S2.1 are satisfied. Consider either a sparse design setting where $2 \leq n_i \leq N_0 < \infty$ or a dense design where $n_i = m \to \infty$, $i = 1, \ldots, n$. Setting $a_n = a_{n1}$ and $b_n = b_{n1}$ for the sparse case, and $a_n = a_{n2}$ and $b_n = b_{n2}$ for the dense case, for a new independent subject i^* and $k \geq 1$, if $m^*(a_n + b_n) = o(1)$ as $n \to \infty$, where $m^* = m^*(n) \to \infty$,*

$$|\hat{\xi}_k^* - \xi_k^*| = O_p(m^{*-1/2} + m^*(a_n + b_n)).$$

This result concerns the transition of sparse to dense sampling specifically for a new subject. A related result was obtained previously in Dai *et al.* (2018) in the Gaussian case. The present result is more general as it does not require Gaussian or any other distributional assumptions.

Write for two sequences θ_n and γ_n that $\theta_n \asymp \gamma_n$ whenever $c_1\theta_n \leq \gamma_n \leq c_2\theta_n$ holds for some constants $c_1, c_2 > 0$ as $n \to \infty$. For dense designs, if the number of individual observations

$m = m(n)$ satisfies $m \asymp (n/\log n)^q$ for some $q \in [1/4, \infty)$, $h_\mu \asymp (\log n/n)^{1/4}$, $h_G \asymp (\log n/n)^\rho$ with $\rho \in (0, 1/4)$, α defined in (A6) satisfies $\alpha > 4$, β_γ defined in (A8) is such that $\beta_\gamma > 2/(1-4\rho)$, where the assumptions are introduced in the Supplement Section S2.1, then $a_n + b_n \asymp (\log n/n)^{2\rho}$.

A larger value of $\rho \in (0, 1/4)$ along with the existence of a suitable $\beta_\gamma = \beta_\gamma(\rho)$ as before leads to a rate $a_n + b_n$ closer to $(\log n/n)^{1/2}$. Here the choice $0 < \rho < 1/4$, which entails the rate for the covariance smoothing bandwidth h_G, is required in order to satisfy condition (A8). If $m^* \asymp (a_n + b_n)^{-\rho_1}$ for some $\rho_1 \in (0, 1)$, the condition $m^*(a_n + b_n) = o(1)$ is satisfied and the rate in Theorem 1 becomes $O_p((\log n/n)^{\rho_1 \rho} + (\log n/n)^{2\rho(1-\rho_1)})$. Hence, larger values of $\rho \in (0, 1/4)$ along with the optimal choice $\rho_1 = 2/3$ lead to an optimal rate arbitrarily close to $O_p((\log n/n)^{1/6})$.

For sparse designs, choosing bandwidths $h_\mu \asymp (\log n/n)^{1/5}$ and $h_G \asymp (\log n/n)^{1/6}$ leads to $a_n + b_n \asymp (\log n/n)^{1/3}$. Taking $m^* \asymp (a_n + b_n)^{-\rho_1}$ for some $\rho_1 \in (0, 1)$, the condition $m^*(a_n + b_n) = o(1)$ is satisfied and the rate in Theorem 1 becomes $O_p((\log n/n)^{\rho_1/6} + (\log n/n)^{(1-\rho_1)/3})$. The optimal rate then becomes $O_p((\log n/n)^{1/9})$, which is achieved when $\rho_1 = 2/3$.

3. Predictive Distributions for Gaussian Processes

Using Gaussianity, for any positive integer K, $\boldsymbol{\xi}_{iK} = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{iK})^T \sim N(0, \boldsymbol{\Lambda}_K)$, where as above $\boldsymbol{\Lambda}_K = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ and $\lambda_k = E(\xi_{ik}^2)$. Conditional on $\mathbf{T}_i$, it follows that $\boldsymbol{\xi}_{iK}$ and $\mathbf{X}_i$ are jointly normal

$$\begin{pmatrix} \mathbf{X}_i \\ \boldsymbol{\xi}_{iK} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_i \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_i & \boldsymbol{\Phi}_{iK}\boldsymbol{\Lambda}_K \\ \boldsymbol{\Lambda}_K\boldsymbol{\Phi}_{iK}^T & \boldsymbol{\Lambda}_K \end{pmatrix} \right).$$

By a well-known property of multivariate normal distributions (see e.g. Mardia *et al.* (1979)),

$$\boldsymbol{\xi}_{iK} | \mathbf{X}_i, \mathbf{T}_i \sim N_K(\tilde{\boldsymbol{\xi}}_{iK}, \boldsymbol{\Sigma}_{iK}), \tag{3.7}$$

where $\tilde{\boldsymbol{\xi}}_{iK} = E(\boldsymbol{\xi}_{iK}|\mathbf{X}_i, \mathbf{T}_i)$ given in (2.4) is the best linear unbiased predictor of $\boldsymbol{\xi}_{iK}$ and $\boldsymbol{\Sigma}_{iK} = \boldsymbol{\Lambda}_K - \boldsymbol{\Lambda}_K \boldsymbol{\Phi}_{iK}^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Phi}_{iK} \boldsymbol{\Lambda}_K$ is the conditional variance. The relation in (3.7) was previously exploited, for example in Yao *et al.* (2005a), to construct simultaneous confidence bands for estimated trajectories; compare also Wang and Shi (2014). We refer to the conditional distribution in (3.7) as *K-truncated predictive distribution* since it is a distributional representation for the subject's truncated true but unobserved scores $\boldsymbol{\xi}_{iK}$.

Note that (2.5) implies that the center of the K-truncated predictive distribution converges to the true FPCs $\boldsymbol{\xi}_{iK}$ as the design gets denser, i.e. as $m \to \infty$. Next, it will be shown that the entire K-truncated predictive distribution shrinks to a point mass located at its true K-truncated FPCs. Recall that $\boldsymbol{\Sigma}_{iK}$ is the conditional covariance as in (3.7) and for a matrix $A \in \mathbb{R}^{p\times q}$ denote by $\|A\|_{\text{op},2} = \sup_{\|v\|_2=1} \|Av\|_2$ the 2-matrix norm, where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^p$, $p, q > 0$. For the following, Gaussianity will be required, i.e.,

(X4) The process $X(t)$, $t \in \mathcal{T}$, and the measurement errors are jointly Gaussian.

Proposition 2. *Suppose that (X1)–(X4) hold and the number of observations for the ith subject diverges, i.e. $n_i = m \to \infty$, $i = 1, \ldots, n$. Then for any fixed $K \geq 1$*

$$\|\boldsymbol{\Sigma}_{iK}\|_{op,2} = O_p(m^{-1}).$$

Note that Gaussianity is used only to derive the explicit form of the conditional covariance $\boldsymbol{\Sigma}_{iK}$ of the FPCs given the data $(\mathbf{X}_i, \mathbf{T}_i)$. We are not aware of any other results in the literature studying the shrinkage of conditional covariance in the dense sampling case.

If Gaussianity does not hold, using the explicit form of $E(\boldsymbol{\xi}_{iK}|\mathbf{X}_i, \mathbf{T}_i)$ in Section 2 and the relation $\text{var}(\boldsymbol{\xi}_{iK} \mid \mathbf{T}_i) = \boldsymbol{\Lambda}_K$, by a conditioning argument $\boldsymbol{\Sigma}_{iK} := E[\,\text{var}(\boldsymbol{\xi}_{iK}|\mathbf{X}_i, \mathbf{T}_i) \mid \mathbf{T}_i] = \text{var}(\boldsymbol{\xi}_{iK} \mid \mathbf{T}_i) - \text{var}(E(\boldsymbol{\xi}_{iK}|\mathbf{X}_i, \mathbf{T}_i) \mid \mathbf{T}_i)$ share the same definition as in the Gaussian case, and therefore Proposition 2 continues to hold for this $\mathbf{\Sigma}_{iK}$. Propositions 1 and 2 demonstrate that the K-truncated predictive distribution of a given subject shrinks to the true K-truncated FPCs $\boldsymbol{\xi}_{iK}$ at a root-m rate as the number of observations per subject diverges. The size of a K-truncated predictive distribution defined through the covariance norm corresponding to the Gaussian distribution (3.7) implicitly reflects the number of available observations.

To discuss this further, consider an independent densely measured subject i^* as in Section 2. The next result quantifies the shrinkage of the conditional variance corresponding to the K-truncated distribution as the number of observations for the subject i^* increases.

Theorem 2. *Suppose that (X2), (X4), (B1) and (A1)–(A8) in the Supplement Section S2.1 hold. Let $K > 0$ be fixed and consider either a sparse design setting when $n_i \leq N_0 < \infty$ or a dense design when $n_i = m \to \infty$, $i = 1, \ldots, n$. Set $a_n = a_{n1}$ and $b_n = b_{n1}$ for the sparse case, and $a_n = a_{n2}$ and $b_n = b_{n2}$ for the dense design. For a new independent subject i^*, if $m^*(a_n + b_n) = o(1)$ as $n \to \infty$, where $m^* = m^*(n) \to \infty$,*

$$\|\hat{\mathbf{\Sigma}}_K^* - \mathbf{\Sigma}_K^*\|_{op,2} = O_p(a_n + b_n).$$

As outlined in Section 2, the estimated covariance $\hat{\mathbf{\Sigma}}_K^*$ for a new subject i^* and thus its K-truncated predictive distribution can be consistently recovered. The shrinkage effect for predictive distributions from sparse to dense is illustrated in Figure 1. The following theoretical framework is a direct consequence of the theory of square integrable Gaussian processes. For the separable real Hilbert space $\mathcal{H} = L^2(\mathcal{T})$ with inner product $\langle\ ,\ \rangle := \langle\ ,\ \rangle_{L^2(\mathcal{T})}$, a probability measure ν defined over the Borel sets $\mathcal{B}(\mathcal{H})$ is Gaussian if for any $h \in \mathcal{H}^*$, where $\mathcal{H}^*$ denotes the dual space consisting of continuous and linear functionals on $\mathcal{H}$, $\mu \circ h$ is a Gaussian measure on $\mathbb{R}$ (Gelbrich, 1990). Such measures ν are characterized by their mean $m_\nu \in \mathcal{H}$ and covariance operator $\Xi_\nu : \mathcal{H} \to \mathcal{H}$ (Kuo,

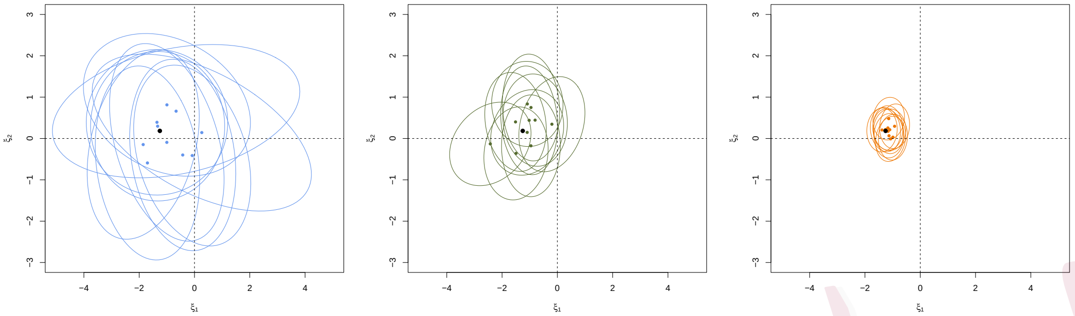

Figure 1: The 95% contours for 10 predictive distributions for the joint distribution of the first two functional principal components with $K = 2$ obtained by random sampling of the data of a new subject when varying the number of observations n_i per subject in the transition from sparse to dense, for $n_i = 2$ (very sparse; left panel), $n_i = 10$ (medium sparse; middle panel), and $n_i = 50$ (dense; right panel), for error variance $\sigma = 0.5^2$ and eigenfunctions $\phi_1(t) = -\cos(\pi t/10)/\sqrt{5}$, $\phi_2(t) = \sin(\pi t/10)/\sqrt{5}$, $\mu(t) = t + \sin(t)$, $t \in \mathcal{T} = [0, 10]$. The time points are sampled from a uniform distribution on $\mathcal{T}$. As expected, the predictive distributions shrink towards a point mass located at the true unobserved functional principal components (black dot) as the data gets denser. The colored dots correspond to the centers of the simulated predictive distributions.

1975), defined through

$$\langle m_\nu, a\rangle = \int_{\mathcal{H}} \langle x, a\rangle \nu(dx), \quad a \in \mathcal{H},$$

$$\langle \Xi_\nu(a), b\rangle = \int_{\mathcal{H}} \langle x - m_\nu, a\rangle \langle x - m_\nu, b\rangle \nu(dx), \quad a, b \in \mathcal{H}.$$

Denote the Gaussian measure ν (depending on the context, in $\mathbb{R}^p$ or L^2) by $\mathbb{G}(m_\nu, \Xi_\nu)$. The K-truncated predictive distribution of the centered process $X_i^c(\cdot)$ given $(\mathbf{X}_i, \mathbf{T}_i)$ is defined as

$$\mathcal{G}_{iK} = (\text{The conditional distribution of } \boldsymbol{\xi}_{iK}^T \boldsymbol{\Phi}_K \mid \mathbf{X}_i, \mathbf{T}_i) = \mathbb{G}(\tilde{\mu}_{iK}, \Xi_{iK}),$$

where $\tilde{\mu}_{iK} = \tilde{\boldsymbol{\xi}}_{iK}^T \boldsymbol{\Phi}_K$, $\boldsymbol{\Phi}_K = (\phi_1, \dots, \phi_K)^T$ are the first K eigenfunctions, and $\Xi_{iK} : L^2(\mathcal{T}) \to L^2(\mathcal{T})$ is the integral operator associated with the covariance function $\Gamma_{iK}(s,t) := \sum_{1 \le j,l \le K} [\boldsymbol{\Sigma}_{iK}]_{jl} \phi_j(s) \phi_l(t)$, with $[\mathbf{A}]_{ij}$ denoting the (i, j)th entry of a matrix $\mathbf{A}$. This is the functional counterpart of the K-truncated predictive distribution in (3.7). We refer to $\mathcal{G}_{iK}$

as the K-*truncated predictive distribution* of the ith subject's latent trajectory. The K-truncated predictive distribution $\mathcal{G}_{iK}$ approximates the *true infinite-dimensional predictive distribution*,

$$\mathcal{G}_i = (\text{The conditional distribution of } (X - \mu) \mid \mathbf{X}_i, \mathbf{T}_i) = \mathbb{G}(\tilde{\mu}_i, \Xi_i), \tag{3.8}$$

where $\tilde{\mu}_i = \Gamma(\cdot, \mathbf{T}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_i)$, $t \in \mathcal{T}$ and Ξ_i is the integral operator associated with the covariance function $\Gamma_i(s,t) = \Gamma(s,t) - \Gamma(s, \mathbf{T}_i)\boldsymbol{\Sigma}_i^{-1}\Gamma(\mathbf{T}_i, t)$, $s, t \in \mathcal{T}$, under the convention that $\Gamma(s, \mathbf{T}_i)$ and $\Gamma(\mathbf{T}_i, t)$ are row and column vectors containing the evaluations of Γ, respectively.

Studied next is the approximation to the true latent trajectory as the truncation point K increases where estimated versions are obtained by replacing population quantities by their estimates, leading to the estimate $\hat{\mathcal{G}}_{iK} = \mathbb{G}(\hat{\mu}_{iK}, \hat{\Xi}_{iK})$ of the K-truncated predictive distribution $\mathcal{G}_{iK}$. Here $\hat{\mu}_{iK} = \hat{\boldsymbol{\xi}}_{iK}^T \hat{\boldsymbol{\Phi}}_K$, and $\hat{\Xi}_{iK}$ is the integral operator associated with the covariance function $\hat{\Gamma}_{iK}(s,t) \coloneqq \sum_{1 \le j,l \le K} [\hat{\boldsymbol{\Sigma}}_{iK}]_{jl} \hat{\phi}_j(s)\hat{\phi}_l(t)$. The corresponding infinite-dimensional version is

$$\hat{\mathcal{G}}_i = \mathbb{G}(\hat{\mu}_i, \hat{\Xi}_i), \tag{3.9}$$

where $\hat{\mu}_i(t) = \hat{\Gamma}(t, \mathbf{T}_i)\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_i)$, $t \in \mathcal{T}$ and $\hat{\Xi}_i$ is the integral operator with kernel $\hat{\Gamma}_i(s,t) \coloneqq \hat{\Gamma}(s,t) - \hat{\Gamma}(s, \mathbf{T}_i)\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\Gamma}(\mathbf{T}_i, t)$, $s, t \in \mathcal{T}$.

To quantify the discrepancy between estimated and true predictive distributions, we adopt the 2-Wasserstein distance $\mathcal{W}_2$ (Villani, 2003), which for two measures ν and τ is

$$\mathcal{W}_2(\nu, \tau) = \left\{ \inf_{A \sim \nu, B \sim \tau} E(\|A - B\|^2) \right\}^{1/2}, \tag{3.10}$$

where the norm $\|\cdot\|$ is either the Euclidean norm for measures supported on $\mathbb{R}^d$, $d \ge 1$, or L^2-norm for measures on the L^2 space, and the infimum is taken over all pairs of random variables A and B with marginal distribution ν and τ, respectively. The shrinkage of the distributions $\mathcal{G}_{iK}$ towards an atomic point mass measure $\mathcal{A}_{X_i^c}$ located at the unobserved latent centered process X_i^c when the number of observations $n_i = m$ diverges and the truncation point $K = K(m)$ suitably grows with m can then be characterized as follows.

Theorem 3. *Suppose that (X1)–(X4) and (B1) hold. Consider a given subject $i \in \{1, \ldots, n\}$ for which one has m measurements with $m \to \infty$. If $K = K(m) \to \infty$ is chosen such that $\sum_{k=1}^{K} \lambda_k^{-1} \asymp m^{1-\delta}$ for some $\delta \in (1/2, 1)$, then*

$$\mathcal{W}_2^2(\mathcal{G}_{iK}, \mathcal{A}_{X_i^c}) = O_p\left(m^{-(2\delta-1)} + \sum_{k=K(m)+1}^{\infty} \lambda_k \right). \tag{3.11}$$

The expectation that implicitly appears in the definition (3.10) of the 2-Wasserstein distance is taken here conditionally on the data for the ith subject $(\mathbf{X}_i, \mathbf{T}_i)$ and the unobserved latent trajectory X_i^c so that the point mass $\mathcal{A}_{X_i^c}$ is well defined. Therefore an O_p-term rather than an O-term appears in (3.11). Shrinkage of the K-truncated predictive distribution towards the latent centered process is tied to the eigenvalue decay. The rate of convergence in (3.11) can be illustrated under polynomial and exponential eigenvalue decay:

(D1) $\lambda_k = k^{-\alpha_0}$ for a constant $\alpha_0 > 1$ and all $k \geq 1$,

(D2) $\lambda_k = \exp(-\alpha_1 k)$ for a constant $\alpha_1 > 0$ and all $k \geq 1$.

Under polynomial decay (D1), it follows that $\sum_{k=1}^{K} \lambda_k^{-1} \asymp K^{1+\alpha_0}$ and also $\sum_{k=K+1}^{\infty} \lambda_k \asymp K^{1-\alpha_0}$, so the condition in Theorem 3 implies that $K \asymp m^{(1-\delta)/(1+\alpha_0)}$ and the optimal rate in (3.11) is given by $m^{(1-\alpha_0)/(1+3\alpha_0)}$. This is achieved by choosing $\delta = 2\alpha_0/(1+3\alpha_0)$ and $K \asymp m^{1/(1+3\alpha_0)}$. Faster eigenvalue decay rates for larger α_0 are associated with slower growth rates for $K = K(m)$ as δ approaches $2/3$. In this case the optimal rate approaches $m^{-1/3}$, which is slower than $m^{-1/2}$. The latter rate can be achieved for a finite-dimensional process, where $\lambda_k = 0$ for all $k \geq k_0$ and some $k_0 > 0$. Under exponential eigenvalue (D2), the optimal rate in Theorem 3 is again $m^{-1/3}$, which is obtained by selecting $\delta = 2/3$ and $K \asymp \log(m^{1/3})$. Note that the result in Theorem 3 is at the population level and does not involve estimation or the sample size n.

The bound (3.11) utilizes the population level K-truncated predictive distribution $\mathcal{G}_{iK}$, which depends upon unknown quantities that must be estimated in practice, which introduces additional errors. The following result establishes consistency of the estimated K-truncated predictive distribution counterpart $\hat{\mathcal{G}}_K^*$ under this scenario for a new subject as described in Section 2. Let $\gamma_K(p,q) = \sum_{k=1}^K \lambda_k^{-p}\delta_k^{-q}$, where p, q are non-negative integers and δ_k are the eigengaps. As in the previous result, K is allowed to diverge.

Theorem 4. *Suppose that assumptions (X2), (X4), (B1) and (A1)–(A8) in the Supplement Section S2.1 are satisfied. Consider either a sparse design setting with $2 \le n_i \le N_0 < \infty$ or a dense design when $n_i = m \to \infty$, $i = 1, \ldots, n$. Set $a_n = a_{n1}$ and $b_n = b_{n1}$ for the sparse case, and $a_n = a_{n2}$ and $b_n = b_{n2}$ for the dense case. For a new subject i^*, suppose that $m^* = m^*(n) \to \infty$ is such that $m^*(a_n + b_n) = o(1)$ as $n \to \infty$. If $K = K(m^*)$ satisfies $(a_n + b_n)\gamma_K(1/2, 1) = o(1)$, $m^*(a_n + b_n)^2\gamma_K(2, 2) = o(1)$, $m^{*2}(a_n + b_n)^2\gamma_K(2, 0) = o(1)$, $m^{*4}(a_n + b_n)^4\gamma_K(2, 2) = o(1)$, $(a_n + b_n)\gamma_K(2, 1) = o(1)$ as $n \to \infty$, and $\sum_{k=1}^K \lambda_k^{-1} \asymp m^{*(1-\delta)}$ for some $\delta \in (1/2, 1)$, then*

$$\mathcal{W}_2^2(\hat{\mathcal{G}}_K^*, \mathcal{A}_{X^{*c}}) = o_p(1).$$

Under polynomial eigenvalue decay (D1) and taking $m^* = m^*(n) \asymp (a_n + b_n)^{-q}$ for some $q \in (0, 2/3)$, it follows from the proof of Theorem 4 that the optimal rate is $(a_n+b_n)^{q(\alpha_0-1)/(3\alpha_0+1)}$, which is achieved by taking $\delta = 2\alpha_0/(3\alpha_0 + 1) \in (1/2, 1)$ and $K \asymp m^{*(1-\delta)/(1+\alpha_0)}$. Thus the optimal rate can be arbitrarily close to $(a_n + b_n)^{2(\alpha_0-1)/(3(1+3\alpha_0))}$ by assuming faster growth rates of m^* with $q \uparrow 2/3$. Faster eigenvalue decay rates, i.e. larger values of α_0, lead to a rate closer to $(a_n + b_n)^{2/9}$. If the eigenvalues exhibit exponential decay (D2) and $m^* = m^*(n) \asymp (a_n + b_n)^{-q}$,

$q \in (0,1)$, the optimal rate is $(a_n + b_n)^{2/9}$.

4. Predictive Distributions in the Functional Linear Model

The concept of predictive distributions is also sensible in the functional linear model (1.3) when predictor processes are sparsely sampled. Suppose one has an infinite-dimensional Gaussian predictor process $X(t)$, $t \in \mathcal{T}$ with a Euclidean response $Y \in \mathbb{R}$. Utilizing the Karhunen–Loève representation (1.2) of predictor processes X and a representation of the slope function $\beta(t) = \sum_{j=1}^{\infty} \beta_j \phi_j(t)$ in the eigenbasis with $\beta_j = \int_{\mathcal{T}} \beta(t)\phi_j(t)$, $j = 1, 2, \ldots$, leads to

$$E(Y|X^c) = \beta_0 + \sum_{j=1}^{\infty} \xi_j \beta_j =: \eta. \tag{4.12}$$

Here $\beta_0 = E(Y)$ is the intercept and η is the linear predictor with responses $Y = \beta_0 + \sum_{j=1}^{\infty} \xi_j \beta_j + \epsilon_Y$, where $\epsilon_Y \sim N(0, \sigma_Y^2)$ is independent of all other random quantities (Yao *et al.*, 2005b; Cai and Hall, 2006; Balasubramanian *et al.*, 2025).

Predicting the scalar response Y (Hall and Horowitz, 2007) based on a sparsely observed predictor process X is clearly of interest. When shifting the focus from point prediction to predictive distributions of the linear predictor η, instead of targeting the distribution for the observed response Y, which contains the additional error ϵ_Y that is independent of all other random quantities and thus is aleatoric and inherently unpredictable, the focus is on the distribution of the predictable part of the response Y. Consider $\eta_K = \beta_0 + \boldsymbol{\beta}_K^T \boldsymbol{\xi}_K$, the truncated real-valued predictor employing the first K principal components, where in practice K can be chosen by a suitable criterion and $\boldsymbol{\beta}_K = (\beta_1, \ldots, \beta_K)^T$. Thus $\eta = \eta_K + \mathcal{R}_K$, where $\mathcal{R}_K = \sum_{j \geq K+1} \xi_j \beta_j$ corresponds to a term that remains unexplained by η_K. This term decreases asymptotically as $E(\mathcal{R}_K) = 0$ and $\text{Var}(\mathcal{R}_K) = \sum_{j \geq K+1} \lambda_j \beta_j^2 = o(1)$ as K increases, where the latter rate can be further specified and can be made arbitrarily fast under additional assumptions (Hall and Horowitz, 2007).

Since X is a Gaussian process, given $\boldsymbol{\beta}_K$ one obtains predictive distributions

$$\mathcal{P}_{iK} \overset{d}{=} N(\beta_0 + \boldsymbol{\beta}_K^T \tilde{\boldsymbol{\xi}}_{iK}, \boldsymbol{\beta}_K^T \boldsymbol{\Sigma}_{iK} \boldsymbol{\beta}_K) \tag{4.13}$$

as before. According to Theorems 1 and 2, these predictive distributions collapse into a point mass located at the true but unobserved predictable part η_{iK} of the response Y_i in the transition from sparse to dense sampling. To quantify the performance of the predictive distribution $\mathcal{P}_{iK}$ (4.13) in the sparse case, it is sensible to employ the 2-Wasserstein distance between two probability measures ν_1, ν_2, which for multivariate distributions is as previously given in (3.10). For our current purpose the predictive distributions are one-dimensional, and for this case (3.10) greatly simplifies and can be expressed as (Villani, 2003)

$$\mathcal{W}_2^2(\nu_1, \nu_2) = \int_0^1 (Q_1(p) - Q_2(p))^2 dp, \tag{4.14}$$

where $Q_j(p) = \inf\{s \in \mathbb{R} \colon F_j(s) \geq p\}$, $p \in (0,1)$, is the quantile function of ν_j, $j = 1, 2$.

To quantify the discrepancy of this predictive distribution, it makes sense to utilize the average Wasserstein distance between $\mathcal{P}_{iK}$ and the atomic measure $\mathcal{A}_{Y_i}$ located at Y_i. Formally,

$$\mathcal{D}_{nK} := n^{-1} \sum_{i=1}^n \mathcal{W}_2^2(\mathcal{A}_{Y_i}, \mathcal{P}_{iK}) = n^{-1} \sum_{i=1}^n (Y_i - \tilde{\eta}_{iK})^2 + n^{-1} \sum_{i=1}^n \boldsymbol{\beta}_K^T \boldsymbol{\Sigma}_{iK} \boldsymbol{\beta}_K, \tag{4.15}$$

where $\tilde{\eta}_{iK} = E(\eta_{iK} | \mathbf{X}_i) = \beta_0 + \boldsymbol{\beta}_K^T \tilde{\boldsymbol{\xi}}_{iK}$ is the best prediction of the truncated linear predictor. Note that (4.15) follows from (4.14) and similar ideas as in Amari and Matsuda (2021) when computing the Wasserstein distance between the predictive distribution and an atomic measure.

If the number of observations $n_i = m_0 < N_0$ is common across subjects, so that the $\boldsymbol{\Sigma}_{iK}$ form an i.i.d. sequence of random positive definite matrices, the proof of Theorem 5 below shows that

$\mathcal{D}_{nK}$ converges to the population-level Wasserstein discrepancy

$$\mathcal{D}_K = 2\boldsymbol{\beta}_K^T E(\boldsymbol{\Sigma}_{1K})\boldsymbol{\beta}_K + \sigma_Y^2 + \sum_{k\geq K+1}\lambda_k\beta_k^2 - 2\boldsymbol{\beta}_K^T E\Big[\boldsymbol{\Lambda}_K\boldsymbol{\Phi}_{1K}^T\boldsymbol{\Sigma}_1^{-1}\sum_{k\geq K+1}\phi_k(\mathbf{T}_1)\lambda_k\beta_k\Big]. \quad (4.16)$$

The first term in (4.16) reflects both the number of observations and the time locations, where increased values of m_0 lead to smaller $\boldsymbol{\beta}_K^T E(\boldsymbol{\Sigma}_{1K})\boldsymbol{\beta}_K$ and thus lower discrepancies, i.e. increased predictability. Similarly, increased predictor and response noise levels σ^2 and σ_Y^2 are associated with worse predictability. The last two terms come from the unexplained linear predictor part $\mathcal{R}_K$ and become smaller as K increases.

Consider an example with eigenbasis $\phi_k(t) = \sin(k\pi t)/\sqrt{2}$, $t \in \mathcal{T}$. If the Fourier coefficients β_k and eigenvalues λ_k exhibit polynomial decay $|\beta_k| = O(k^{-\alpha_1})$ and $\lambda_k = O(k^{-\alpha_2})$, $\alpha_1, \alpha_2 > 1$, the Cauchy–Schwarz inequality implies $\sum_{k\geq K+1}\lambda_k\beta_k^2 = O(K^{1-2\alpha_1-\alpha_2})$ and similarly $\boldsymbol{\beta}_K^T E[\boldsymbol{\Lambda}_K\boldsymbol{\Phi}_{1K}^T\boldsymbol{\Sigma}_1^{-1}\sum_{k\geq K+1}\phi_k(\mathbf{T}_1)\lambda_k\beta_k] = O(K^{1-\alpha_1-\alpha_2})$ with $K^{1-\alpha_1-\alpha_2} \leq K^{-1}$, where one uses that $\|\boldsymbol{\beta}_K\|_2 \leq \|\beta\|_{L^2}$ and a uniform bound on the remaining quantities; see the proof of Supplement Lemma S12. In practice, the predictive distributions $\mathcal{P}_{iK}$ and therefore also the $\mathcal{D}_{nK}$ are unknown as they depend on unknown population quantities; substituting estimates for these quantities results in estimates $\hat{\mathcal{P}}_{iK}$ and $\hat{\mathcal{D}}_{nK}$. To obtain estimates of β_0 and slope coefficients $\boldsymbol{\beta}_K$, one can adopt a standard approach under the following assumption (B2),

(B2) $\|\beta\|_{L^2}^2 = \sum_{m=1}^{\infty}\sigma_m^2/\lambda_m^2 < \infty$.

With $C(t) = \text{Cov}(X(t), Y) = \sum_{k=1}^{\infty}E(Y\xi_k)\phi_k(t)$ denoting the cross-covariance function between the process X and response Y and $\sigma_k = \int_{\mathcal{T}} C(t)\phi_k(t)dt = E(Y\xi_k)$, $k = 1, 2, \ldots$, one can estimate $C(t)$ using a local linear smoother on the raw covariances $C_i(T_{ij}) = (\tilde{X}_{ij} - \hat{\mu}(T_{ij}))Y_i$ (Yao *et al.*, 2005b), leading to an estimate $\hat{C}(t)$ that depends on a bandwidth h; see Lemma S9 in the Supplementary Material for details. Since $\sigma_k = \lambda_k\beta_k$, under (B2), it holds that $\beta(t) =$

$\sum_{m=1}^{\infty} \sigma_m \phi_m(t)/\lambda_m$, $t \in \mathcal{T}$. This motivates to estimate β by

$$\hat{\beta}_M(t) = \sum_{m=1}^{M} \frac{\hat{\sigma}_m}{\hat{\lambda}_m} \hat{\phi}_m(t), \quad t \in \mathcal{T},$$

where $\hat{\sigma}_k = \int_{\mathcal{T}} \hat{C}(t)\hat{\phi}_k(t)dt$ is an estimate of σ_k and $M = M(n)$ is a positive integer sequence that diverges as $n \to \infty$. The intercept $\beta_0 = E(Y)$ is estimated by $\hat{\beta}_0 = n^{-1}\sum_{i=1}^{n} Y_i$. Convergence of $\hat{\beta}_M$ towards β is tied to the eigengaps of X (Cai and Hall, 2006; Müller and Yao, 2010).

With estimates $\hat{\beta}_M$ of β in hand, one readily obtains estimates of the predictive distributions $\hat{\mathcal{P}}_{iK}$. We assume for simplicity that the optimal asymptotic tuning parameters are used for estimating the mean, covariance and cross-covariance, $h_\mu \asymp (\log n/n)^{1/5}$, $h_G \asymp (\log n/n)^{1/6}$ (Dai *et al.*, 2018) and $h \asymp n^{-1/3}$ in the sparse design situation; in particular, this implies $c_n := \max(a_n, b_n) \asymp (\log n/n)^{1/3}$. Defining sequences $\upsilon_M = \sum_{m=1}^{M} \delta_m^{-1}$, $\tau_M = \sum_{m=1}^{M} \lambda_m^{-1}$ and a remainder term $\Theta_M = \|\sum_{m\geq M+1}(\sigma_m/\lambda_m)\phi_m\|_{L^2}$, where δ_m are the eigengaps, note that $M = M(n)$ should not grow too fast with sample size n :,

(B3) The integer sequence $M = M(n) \to \infty$ as $n \to \infty$ is such that

$\sum_{m=1}^{M} \lambda_m^{-1/2}\delta_m^{-1} = O(c_n^{\rho-1})$ for some $\rho \in (1/3, 1)$,

with an additional regularity assumption to obtain uniform convergence,

(C1) There exists a scalar $\kappa_0 > 0$ such that $\lambda_{\min}(\mathbf{\Sigma}_{iK}) \geq \kappa_0$ almost surely, for all $i \geq 1$.

(C1) is a mild assumption, as $\mathbf{\Sigma}_{iK}$ corresponds to the conditional variance of $\boldsymbol{\xi}_{iK} - \tilde{\boldsymbol{\xi}}_{iK}$ given $\mathbf{T}_i$, which is positive definite and does shrink to zero in the sparse case, due to $n_i \leq N_0 < \infty$.

Our next result demonstrates that $\hat{\mathcal{P}}_{iK}$ is consistent for $\mathcal{P}_{iK}$ in the 2-Wasserstein metric, the Kolmogorov metric and in the L^2 metric. Let F_{iK} denote the cumulative distribution function corresponding to $\mathcal{P}_{iK}$ in (4.13) and $\hat{F}_{iK}$ that obtained by replacing $\tilde{\boldsymbol{\xi}}_{iK}$ and $\mathbf{\Sigma}_{iK}$ by $\hat{\boldsymbol{\xi}}_{iK}$ and $\hat{\mathbf{\Sigma}}_{iK}$, respectively, and β_0 and $\boldsymbol{\beta}_K$ by the above estimates. Denote the estimated and true predictive densities by $\hat{f}_i(t) = d\hat{F}_i(t)/dt$ and $f_i(t) = dF_i(t)/dt$. The L^2 norm of a function $g \colon \mathcal{T} \to \mathbb{R}$ is $\|g\|_{L^2(\mathbb{R})} = (\int_{\mathbb{R}} g^2(s)ds)^{1/2}$.

(B4) Let $c_n = \max(a_n, b_n) \to 0$ as $n \to \infty$, where a_n and b_n are defined in (2.6).

Theorem 5. *Suppose that (X4), (B1)–(B4), (A1)–(A8) in the Supplement Section S2.1 hold, and consider a sparse design with $n_i \leq N_0 < \infty$. For a fixed $K \geq 1$, setting $a_n = a_{n1}$ and $b_n = b_{n1}$,*

$$\mathcal{W}_2(\hat{\mathcal{P}}_{iK}, \mathcal{P}_{iK}) = O_p(\alpha_n), \tag{4.17}$$

$$\sup_{t \in \mathbb{R}} |\hat{F}_{iK}(t) - F_{iK}(t)| = O_p(\alpha_n), \tag{4.18}$$

$$\|\hat{f}_{iK} - f_{iK}\|_{L^2(\mathbb{R})} = O_p(\alpha_n), \tag{4.19}$$

as $n \to \infty$, where $\alpha_n = c_n \upsilon_M + c_n^{\rho} \tau_M^{1/2} + \Theta_M$ and the $O_p(\alpha_n)$ terms are uniform in i.

Under the conditions of Theorem 5, $\alpha_n \to 0$ is a consequence of $\tau_M \leq \upsilon_M = O(c_n^{\rho-1})$, which implies $\alpha_n \leq O(c_n^{(3\rho-1)/2} + \Theta_M)$. There is a trade-off between how fast M can grow and the rate of convergence for the estimates of the population quantities, where a larger M entails a lower remainder term Θ_M but affects the rate at which β is recovered through $\hat{\beta}_M$, which involves M components, and vice versa. Since the former term is connected to the decay of the covariance terms σ_m/λ_m, the optimal growth rate of $M(n)$ is inherently tied to the decay rate of σ_m, λ_m and the eigengaps δ_m.

It is of interest to consider the special case where X is a Brownian motion, for which the λ_m and ϕ_m are well-known (Hsing and Eubank, 2015). Although Brownian motion does not satisfy the smoothness assumptions required, it still provides insight into how the convergence rate is related to the eigenvalue decay of the process. by Lemma S16 in the Supplement, if $M =$

$M(n) \asymp (\log n/n)^{(\rho-1)/15}$, then M satisfies (B3) with $\sum_{m=1}^{M} \lambda_m^{-1/2} \delta_m^{-1} \asymp c_n^{\rho-1}$. Moreover, if the decay of σ_m is such that $\sigma_m^2 \leq Cm^{-(8+\delta)}$ for some constant $C > 0$ and $\delta > 0$, then (B2) is satisfied, the remainder $\Theta_M = O\left(M^{-(1+\delta/2)}\right)$ and the rate α_n satisfies the following conditions as stated in Lemma S16: If $\rho \leq (5+\delta)/(15+\delta)$, then $\alpha_n = O((\log n/n)^{(13\rho-3)/30})$ while if $\rho > (5+\delta)/(15+\delta)$ it holds that $\alpha_n = O((\log n/n)^{(1-\rho)(1+\delta/2)/15})$. The optimal rate is achieved when $\rho = (5+\delta)/(15+\delta)$ and leads to $\alpha_n = O((\log n/n)^q)$, where $q = ((2+\delta)/(15+\delta))/3$. A sufficiently large δ implies that q is closer to $1/3$ so that the rate α_n approaches $c_n = (\log n/n)^{1/3}$, which is the rate at which population quantities such as the covariance function Γ are uniformly recovered (see e.g. Theorem 5.2 in Zhang and Wang (2016)).

Regarding the Wasserstein discrepancy $\mathcal{D}_{nK}$, the proposed predictability measure and the response measurement error variance σ_Y^2 can be consistently estimated in the sparse case. Consider the special case where the number of observations $n_i = m_0 < N_0$ is common across subjects. Then the estimated Wasserstein discrepancy measure $\hat{\mathcal{D}}_{nK}$ converges to the population target $\mathcal{D}_K$.

Theorem 6. *Suppose that (X4), (B1)–(B4), (C1), (A1)–(A8) in the Supplement Section S2.1 hold and consider a sparse design with $n_i = m_0 \leq N_0 < \infty$, setting $a_n = a_{n1}$ and $b_n = b_{n1}$. For $K \geq 1$,*

$$\hat{\mathcal{D}}_{nK} = \mathcal{D}_K + O_p(\alpha_n), \quad \alpha_n = c_n \upsilon_M + c_n^{\rho} \tau_M^{1/2} + \Theta_M, \tag{4.20}$$

and furthermore

$$n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 - \sum_{m=1}^{M} \hat{\lambda}_j \hat{\beta}_j^2 = \sigma_Y^2 + O_p(\alpha_n) + \sum_{m \geq M+1} \lambda_m \beta_m^2 \tag{4.21}$$

with $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$.

Of interest is also the behavior of the estimated predictive distributions under the transition from sparse to dense sampling for a new independent subject i^*.

Theorem 7. *Suppose that assumptions (X2), (X4), (B1)–(B4) and (A1)–(A8) in the Supplement Section S2 are satisfied. Consider either a sparse design setting when* $2 \leq n_i \leq N_0 < \infty$ *or a dense design when* $n_i = m \to \infty$, $i = 1, \ldots, n$. *Set* $a_n = a_{n1}$ *and* $b_n = b_{n1}$ *for the sparse case and* $a_n = a_{n2}$ *and* $b_n = b_{n2}$ *for the dense case. Let* $K > 0$ *be fixed and take* $h = n^{-1/3}$. *For a new independent subject* i^*, *suppose that* $m^* = m^*(n) \to \infty$ *is such that* $m^*(a_n + b_n) = o(1)$ *as* $n \to \infty$. *Then*

$$\mathcal{W}_2^2(\mathcal{P}_K^*, \mathcal{A}_{\beta_0 + \boldsymbol{\beta}_K^T \boldsymbol{\xi}_K^*}) = O_p(m^{*-1}),$$

$$\mathcal{W}_2^2(\hat{\mathcal{P}}_K^*, \mathcal{A}_{\beta_0 + \boldsymbol{\beta}_K^T \boldsymbol{\xi}_K^*}) = O_p\left(m^{*2}(a_n + b_n)^2 + m^{*-1} + a_n + b_n + r_n^{*2}\right),$$

where $r_n^* = c_n \upsilon_M + c_n^\rho \tau_M^{1/2} + \tau_M \left[n^{-1/3} + a_n \right] + \Theta_M$.

5. Simulations

To illustrate the theoretical results in Propositions 1 and 2 pertaining to the convergence of the best linear unbiased predictors $\tilde{\xi}_{ik}$ to the FPC scores ξ_{ik} and the shrinkage of the conditional variance $\boldsymbol{\Sigma}_{iK}$ in the transition from sparse to dense sampling designs, consider a finite-dimensional Gaussian process $X(t)$, $t \in \mathcal{T} = [0, 10]$, generated from four principal components with population quantities given by $\phi_1(t) = -\cos(\pi t/10)/\sqrt{5}$, $\phi_k(t) = \sin((2k-3)\pi t/10)/\sqrt{5}$, $k = 2, \ldots, 4$, $\mu(t) = t + \sin(t)$, $\lambda_k = 5 - k$, $k = 1, \ldots, 4$, and $\sigma = 0.5$. It is of interest to consider a range of sampling designs including very sparse ($n_i = m = 2$), medium sparse ($n_i = m = 10$), and dense ($n_i = m = 50$) designs. The time points are selected at random without replacement from an equispaced grid of $2,000$ points over $\mathcal{T}$.

Figure 2 shows the boxplot for $||\tilde{\boldsymbol{\xi}}_{iK} - \boldsymbol{\xi}_{iK}||_2$ and $||\boldsymbol{\Sigma}_{iK}||_{\text{op},2}$ in the transition from sparse to dense sampling across 200 simulations and for a truncation parameter $K = 2$. Clearly, both errors terms shrink towards zero as the sampling design gets denser, indicating the convergence of each $\tilde{\xi}_{ik}$ to the FPC score ξ_{ik}, $k = 1, 2$, since $|\tilde{\xi}_{ik} - \xi_{ik}| \leq ||\tilde{\boldsymbol{\xi}}_{iK} - \boldsymbol{\xi}_{iK}||_2$, but also the shrinkage of the entire conditional distribution. To demonstrate how the finite sample results conform with the theory for the FLM model, consider the following population quantities: $\mu(t) = t/2$, $\lambda_k = 4/(1+k)^2$, $k = 1, \ldots, 4$, $K = 4$, and the intercept and slope coefficients $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = -1$, $\beta_3 = 0.5$ and $\beta_4 = -0.5$. We investigate various noise levels for the predictor process X and response Y as well as a variety of sparse settings, where we generate $n_i = m_0$ random time points for the ith subject, $i = 1, \ldots, n$. Here $m_0 = 2$ reflects a very sparse design, $m_0 = 8$ a medium sparse and $m_0 = 20$ a dense design. We then select the time points at random and without replacement from an equi-spaced grid of 100 points over $\mathcal{T}$. Finally, we performed $2,000$ simulations with Julia, interfacing with R and the fdapace package (Gajardo *et al.*, 2021).

Table 1 presents the results for the Wasserstein discrepancy $\hat{\mathcal{D}}_{nK}$ under different sparsity designs and noise levels in both the functional predictor and scalar response Y. The discrepancy $\hat{\mathcal{D}}_{nK}$ reflects the improvements in predictability for lower noise levels and under increasingly denser designs and increases monotonically in both σ and σ_Y and decreases monotonically as the design becomes denser when keeping the noise level σ and σ_Y fixed. As an additional measure of performance for $\mathcal{P}_{iK}$, we computed the estimated 2-Wasserstein distance between the empirical distribution of $\hat{F}_{iK}(\beta_0 + \int_{\mathcal{T}} \beta(s)(X_i(s) - \mu(s))ds)$, $i = 1, \ldots, n$ and a uniform distribution on $(0, 1)$. Further results and discussion can be found in the Supplement Section S1.

Figure 3 displays boxplots for the true underlying discrepancy measure $\mathcal{D}_{nK}$ as the sampling design gets denser for different noise levels in the predictor process and response, which further demonstrates the improvement in predictability for lower noise levels and denser sampling designs as observed before for the estimated discrepancy measure. In addition, Figure 4 illustrates predictive distributions for sparse to dense sampling design scenarios for the underlying predictor process X for noise level $\sigma = \sigma_Y = 0.5$. Clearly, as the sampling design becomes less sparse, the predictive distributions shrink towards the predictable part of the response.

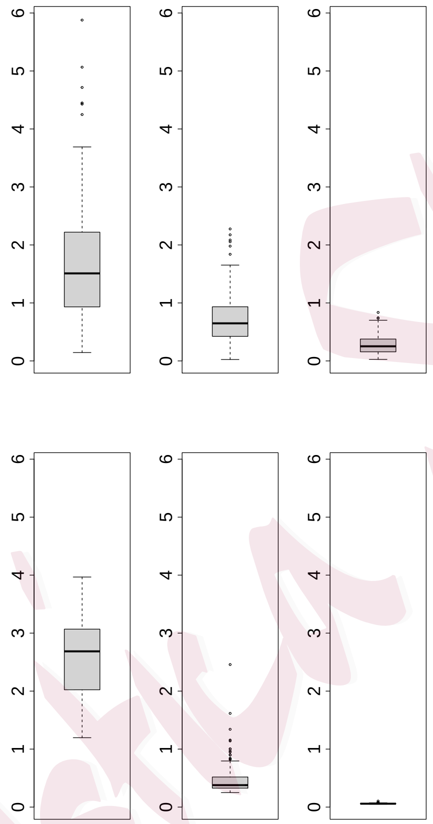

Figure 2: Simulation results illustrating Propositions 1 and 2 with $K = 2$. The upper panel shows boxplots across 200 simulations of the error term $||\tilde{\boldsymbol{\xi}}_{iK} - \boldsymbol{\xi}_{iK}||_2$ for very sparse ($m = 2$, left), less sparse ($m = 10$, middle) and more dense ($m = 50$, right) designs. The lower panel shows the corresponding results for $||\boldsymbol{\Sigma}_{iK}||_{\text{op},2}$.

6. Data Illustration

The concept of predictive distributions for longitudinal data in the context of functional linear regression models is demonstrated for the body mass index (BMI) and systolic blood pressure

Table 1: Averages of the Wasserstein discrepancy $\hat{\mathcal{D}}_{nK}$ (4.15), which measures the predictability of the responses Y_i in the functional linear model by the predictive distribution $\mathcal{P}_{iK}$, obtained for 2000 simulation runs. The true regression parameters are $\beta_0 = 0.5$ and $\boldsymbol{\beta}_K = (1, -1, 0.5, -0.5)^T$. Results are for various predictor and response measurement error and sparsity levels. The very sparse case corresponds to $m = 2$, the sparse case to $m = 8$ and the less sparse design case to $m = 20$ observations per subject.

Measurement Error Noise level		Sparsity setting					
Predictor	Response	Very Sparse		Sparse		Less Sparse	
σ	σ_Y	$n = 500$	$n = 2000$	$n = 500$	$n = 2000$	$n = 500$	$n = 2000$
0.5	0.5	3.008	2.645	1.492	1.477	0.863	0.853
	1.0	3.863	3.421	2.255	2.237	1.612	1.606
1.0	0.5	3.639	3.449	2.540	2.418	1.729	1.715

(SBP) data in the Baltimore Longitudinal Study of Aging (BLSA, Shock *et al.*, 1984), with sparse longitudinal measurements for each subject. This dataset has been analyzed previously in Yao *et al.* (2005b), where one can find further details. We consider a sample of 713 male subjects aged between 50 and 80 years for which their SBP and BMI measurements are within the corresponding 1% and 99% quantiles across all subjects. For the estimation of population quantities the fdapace R package (Gajardo *et al.*, 2021) was used and estimated predictive distributions were constructed as described in Section 4, regressing SBP (in mm Hg) at the last age where it is measured as scalar response against the sparsely observed functional predictor (BMI in kg/m^2). We utilize the first $K = 3$ functional principal component scores of the BMI trajectory, which explain more than 98% of the variation, and choose $M = K$ components and the cross-covariance bandwidth h by leave-one-out cross-validation.

The estimated eigenfunctions are in Figure 5. They reflect the modes of variation in the sample of functional data, where the bandwidths used for the mean and covariance smoothing steps are 1.3 and 2.6, respectively. The first eigenfunction reflects a variation in the overall BMI base level across all ages, whereas the higher order eigenfunctions reflect different BMI contrasts between younger and older ages. For example, the second eigenfunction reflects a mode of variation that differentiates higher BMI levels at ages below 62 years old from lower BMI levels afterwards.

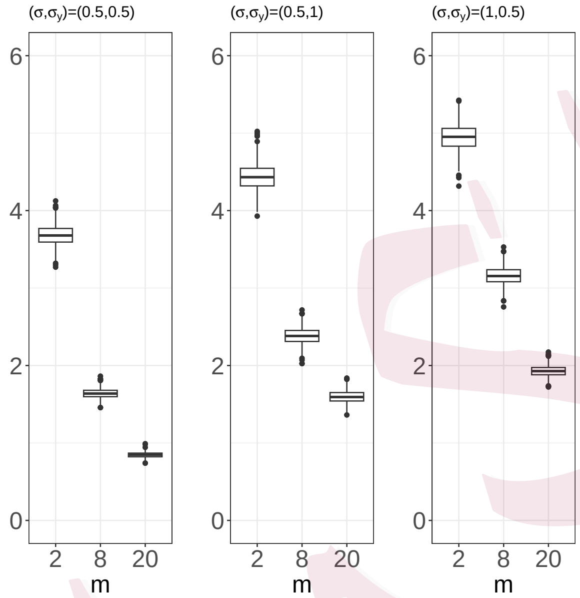

Figure 3: Boxplots of the true underlying Wasserstein discrepancy measure $\mathcal{D}_{nK}$ (4.15) in the functional linear model for 1000 simulations and sample size $n = 500$, for increasingly less sparse sampling designs and various noise levels for the predictor process X and response Y.

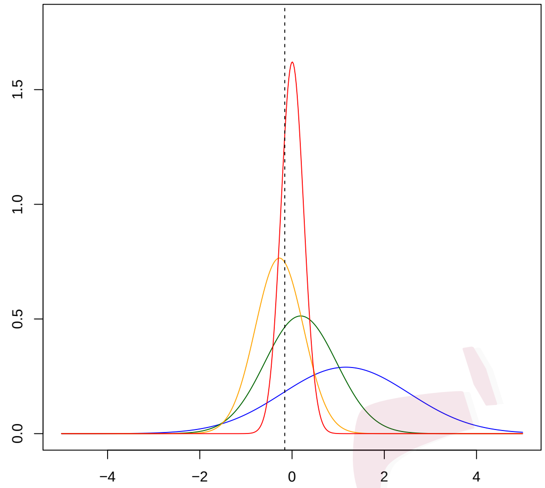

Figure 4: Predictive distributions $\mathcal{P}_K$ for the response in the functional linear model obtained by simulating different sampling design scenarios for a given realization of the predictor process X, for very sparse $m = 2$ (blue), sparse $m = 8$ (green), less sparse $m = 20$ (orange) and dense design $m = 100$ (red), with $\sigma = \sigma_Y = 0.5$. The vertical line corresponds to the (unobserved) predictable part η_K of the response.

Figure 6 illustrates predictive distribution intervals constructed from the 5% and 95% quantiles of the predictive distribution $\hat{\mathcal{P}}_{iK}$ for 20 subjects, where we order them from lowest to largest mean of the predictive distribution. Here it is necessary to emphasize that these intervals are for the prediction intervals for $E(Y|X)$ in the functional linear model and not for the responses Y, which for SBP are known to have a large variance, which means that $E(Y|X)$ will usually be far from Y.

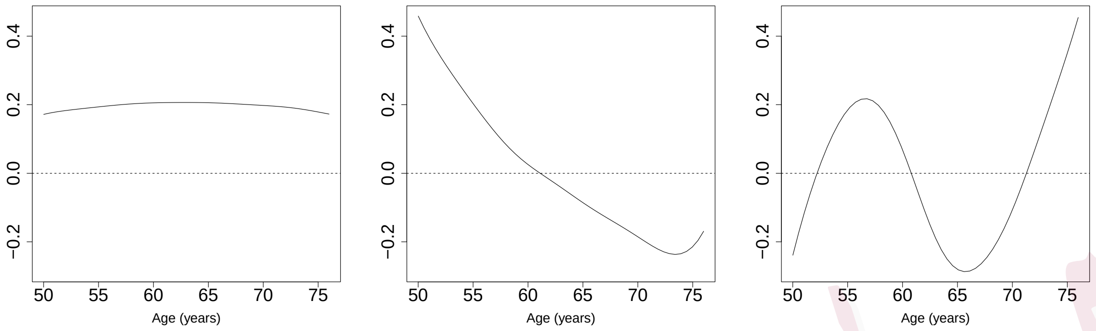

Figure 5: The first three estimated eigenfunctions reflecting the main modes of variation in the sample of sparsely observed BMI functional data from the Baltimore Longitudinal Study of Aging.

7. Discussion and Concluding Remarks

The main message of our paper concerns the common scenario of sparsely observed functional data, which covers many longitudinal designs, where one has only few time points at which noisy measurements of the function are available. In these situations a point estimation perspective is not productive for the functional principal components that figure in the Karhunen–Loève expansion of FPCA since consistent estimators are unavailable. We advocate not to target point estimates of functional principal components and responses given sparse measurements of functional predictors but instead to target predictive distributions, for which consistent estimators are available, leading to prediction regions as the targets of interest.

The inherent uncertainty caused by the sparsity of the measurement times is even present for more densely sampled functional data but to a lesser extent and may then be ignored. This paper provides a formal analysis and precise characterization of the decline in uncertainty as designs get denser. The increasing information content in a design as it gets denser is accurately reflected in the shrinkage of the conditional distributions as delineated in Propositions 1 and 2.

When one aims at a response in a functional linear regression model the predictive distribution targets the (truncated) predictable part of the response Y_i, which is the part that is not contaminated with unpredictable measurement error ϵ_{iY}. Therefore, the observed Y_i, which includes measurement error, is not necessarily located within the prediction region constructed from $\mathcal{P}_{iK}$. Instead, the predictive intervals target the true truncated predictable part $\eta_{iK} = \beta_0 + \boldsymbol{\beta}_K^T \boldsymbol{\xi}_{iK}$ of the observed response Y_i, which is close to the linear predictor $\eta_i = \beta_0 + \sum_{k=1}^{\infty} \beta_k \xi_{ik}$ for a large enough truncation point K.

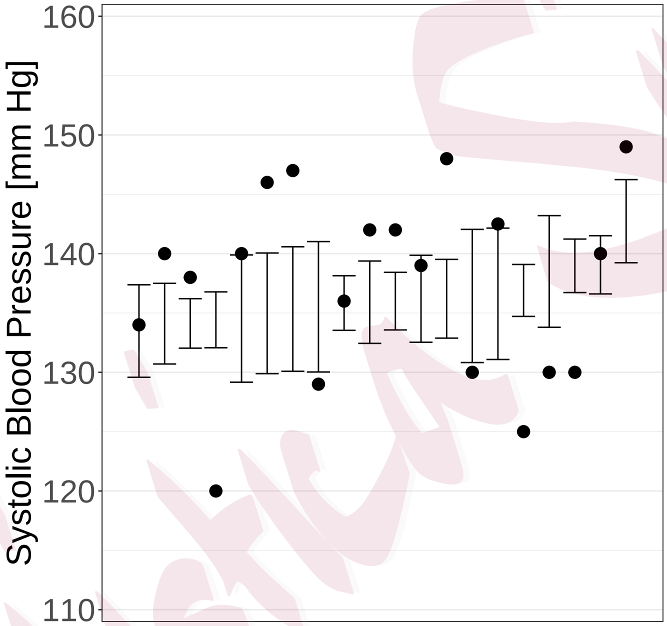

Figure 6: Predictive distribution intervals for $E(Y|X)$ where X are sparsely observed BMI trajectories and Y is the last observed systolic blood pressure. The intervals are ordered from left to right by size of the mean of the predictive distribution, where the interval for the smallest mean is at the left end and the interval with the largest mean at the right end. The dots are the observed responses Y which carry a large random component that is unpredictable.

From a practical perspective, the main implication of the predictive distribution approach that we advocate here is to abandon inconsistent point estimates of functional principal components and their associated trajectories and of predicted responses in the presence of sparsely sampled functional predictors. Instead one should focus on obtaining and using predictive distributions. Since under Gaussian assumptions consistent estimation of these predictive distributions is feasible and theoretically supported with convergence rates, this approach provides for valid uncertainty quantification of functional trajectories and the predictable part of the response whenever predictors are sparsely sampled, as is common in longitudinal designs. Obtaining these predictive distributions is straightforward and can also be used to simulate the effects of different sampling schemes on the uncertainty of the resulting prediction of trajectories and responses.

Supplementary Materials

Additional simulation results, proofs, and auxiliary results are available in the Supplement.

Acknowledgments

The research of XD has been supported by NSF grant DMS-2329879 and the research of HGM by NSF grant DMS-2310450. We express our thanks to the reviewers for helpful comments that led to numerous improvements.

References

Amari, S.-i. and Matsuda, T. (2021). Wasserstein statistics in one-dimensional location scale models. *Annals of the Institute of Statistical Mathematics*.

Balasubramanian, K., Müller, H.-G. and Sriperumbudur, B. K. (2025). Functional linear and single-index models: A unified approach via gaussian stein identity. *Bernoulli* **31**, 973–1006.

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics* **34**, 2159–2179.

Castro, P. E., Lawton, W. H. and Sylvestre, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28**, 329–337.

Chiou, J.-M., Yang, Y.-F. and Chen, Y.-T. (2016). Multivariate functional linear regression and prediction. *Journal of Multivariate Analysis* **146**, 301–312.

Dai, X., Müller, H.-G. and Tao, W. (2018). Derivative principal component analysis for representing the time dynamics of longitudinal and functional data. *Statistica Sinica* **28**, 1583–1609.

Gajardo, A., Carroll, C., Chen, Y., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., Ji, H., Müller, H.-G. and Wang, J.-L. (2021). *fdapace: Functional Data Analysis and Empirical Dynamics*. URL `https://CRAN.R-project.org/package=fdapace`. R package version 0.5.7.

Gelbrich, M. (1990). On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten* **147**, 185–203.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics* **35**, 70–91.

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

Horvath, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. New York: Springer.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons.

Kleffe, J. (1973). Principal components of random variables with values in a separable Hilbert space. *Mathematische Operationsforschung und Statistik* **4**, 391–406.

Kneip, A., Poss, D. and Sarda, P. (2016). Functional linear regression with points of impact. *Annals of Statistics* **44**, 1–30.

Kuo, H.-H. (1975). *Gaussian Measures in Banach spaces*. Springer.

Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics* **38**, 3321–3351.

Mardia, K., Kent, J. and Bibby, J. (1979). *Multivariate Analysis*. Academic Press.

Müller, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics* **32**, 223–240.

Müller, H.-G. and Yao, F. (2010). Empirical dynamics for longitudinal data. *Annals of Statistics* **38**, 3458 – 3486.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. New York: Springer. second edn.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–259.

Shen, X. (2002). Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association* **97**, 222–235.

Shi, J. Q. and Choi, T. (2011). *Gaussian Process Regression Analysis for Functional Data*. CRC Press.

Shock, N. W., Greulich, R. C., Andres, R., Lakatta, E. G., Arenberg, D. and Tobin, J. D. (1984). Normal human aging: The Baltimore longitudinal study of aging. In *NIH Publication No. 84-2450*. Washington, D.C.: U.S. Government Printing Office.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

Wang, B. and Shi, J. Q. (2014). Generalized gaussian process regression model for non-gaussian functional data. *Journal of the American Statistical Association* **109**, 1123–1133.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application* **3**, 257–295.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

— (2005b). Functional linear regression analysis for longitudinal data. *Annals of Statistics* **33**, 2873 – 2903.

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *Annals of Statistics* **44**, 2281–2321.

— (2018). Optimal weighting schemes for longitudinal and functional data. *Statistics & Probability Letters* **138**, 165–170.