Electronic Journal of Statistics
Vol. 15 (2021) 3996–4026
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1882

Shape-preserving prediction for stationary functional time series

Shuhao Jiao* **and Hernando Ombao**

King Abdullah University of Science and Technology,
Thuwal, Saudi Arabia, 23955
e-mail: shuhao.jiao@kaust.edu.sa; hernando.ombao@kaust.edu.sa

Abstract: This article presents a novel method for prediction of stationary functional time series, in particular for trajectories that share a similar pattern but display variable phases. The limitation of most of the existing prediction methodologies for functional time series is that they only consider vertical variation (amplitude, scale, or vertical shift). To overcome this limitation, we develop a shape-preserving (SP) prediction method that incorporates both vertical and horizontal variation. One major advantage of our proposed method is the ability to preserve the shape of functions. Moreover, our proposed SP method does not involve unnatural transformations and can be easily implemented using existing software packages. The utility of the SP method is demonstrated in the analysis of non-metanic hydrocarbons (NMHC) concentration. The analysis demonstrates that the prediction by the SP method captures the common pattern better than the existing prediction methods and also provides competitive prediction accuracy.

MSC2020 subject classifications: 62R10, 37M10.
Keywords and phrases: Functional registration, functional time series, (spherical)K-means clustering, nonlinear dimension reduction, prediction, shape space, state-space model.

Received November 2020.

Contents

*Corresponding author.

1. Introduction

When continuous-time records are separated into natural consecutive time intervals, such as days, weeks, or years, for which a reasonably similar behavior is expected, the resulting functions can be described as a functional time series, where one unit of observation is an observed trajectory. Functional data sometimes exhibit two types of variation: amplitude variation which corresponds to the size or scale of trajectory features, and phase variation which accounts for location variation and temporal shifts. In this paper, we analyzed the curves of non-metanic hydrocarbons (NMHC) collected at road level in an Italian city. In Figure 1, 7 consecutive trajectories of NMHC concentration are displayed. Observe that each daily curve has two peaks. The time of the occurrence of peaks varies across days due to human activity or some other environmental reasons. The variation of the occurrence time of the peaks can be viewed as phase variation. However, existing work typically consider only the vertical variation (i.e., amplitude), but not the variation in phase. An immediate result is that, the predicted curve may not show the common underlying pattern. To overcome this serious limitation, we develop a novel method for stationary functional time series, where trajectories share a common pattern. Our goal is not only to obtain competitive prediction from the past data by some stationary functional time series model, such as functional auto-regressive model, in terms of mean squared error, but also to preserve the underlying pattern for the predicted curves.

There are available prediction methods for stationary functional time series. Besse et al. [3] proposed a non-parametric kernel predictor. Antoniadis and Sapatinas [1] studied the first-order functional autoregression curve prediction based on a linear wavelet method. Kargin and Onaski [18] introduced the predictive factor method. Aue et al. [2] developed a method that uses multivariate techniques in functional time series prediction. Jiao et al. [16] proposed a partial functional prediction method, for the cases where the functions to be

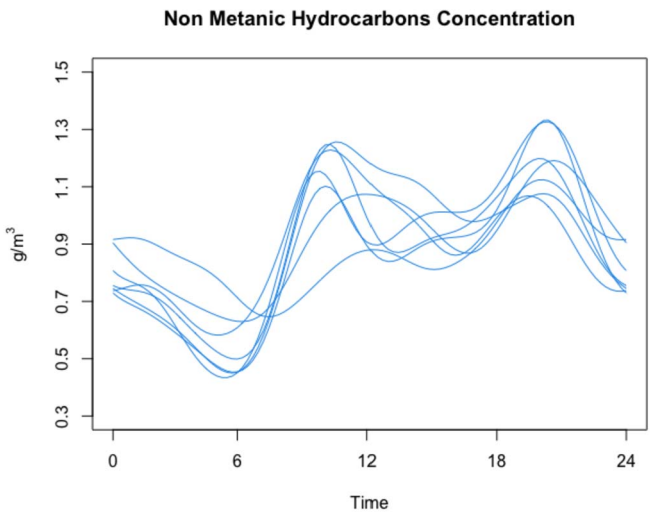

FIG 1. *7 consecutive NMHC concentration trajectories of one week*

predicted are partially observed. There are also some other prediction methods for functional time series, and these methods have their own advantages. However, their main limitation is that they incorporate only vertical variation among the curves. Hyndman and Shang [11] proposed to use weighted functional principal component regression and weighted functional partial least squares regression. One attractive feature of their method is its ability to take account for changes in shape over time. However, the geometrically decaying weights restrict the shape of each function to be close to the neighboring functions. Hence, while this method works well for processes with slowly-evolving shape, its disadvantage is that it is not suitable for situations where "neighboring" functions have different shape or shape changes quickly across curves. Compared to Hydman's method, the SP method allows fast transition of phase. Some related work assume that functions are composed of multiple components which repeat themselves over different periods of time (see e.g., Lin et al. [12] and Lin et al. [13]). The difference between their work and our proposed method is that we assume that there is only one common pattern that repeats itself over the same period of time across curves. This, we believe, is more suitable for some cases such as the environmental data that is being analyzed in this paper.

When trajectories share a common pattern and meanwhile present phase variation, a typical technique researchers usually adopt is functional registration. In functional registration, each function is decomposed as $f_n(t) = X_n \circ \gamma_n(t)$, where the amplitude function $X_n(t)$ accounts for the vertical variation, and the warping function $\gamma_n(t)$ captures the phase information. However, to the best of our knowledge, methods for functional time series prediction have not incorporated functional registration. The prediction method that we develop here involves the prediction of amplitude functions and warping functions. One

of the major challenges is the prediction of warping functions, since they do not lie in a linear space, and thus ordinary linear models are not applicable. Warping functions must be monotonically increasing, and are restricted to start and end at two fixed values. There are several ways to model warping functions. Generally speaking, all these methods seek to apply linear models to non-linear objects.

It is noted that warping functions share similar properties with probability distribution functions, that is, they are all non-decreasing and have common starting and ending values. There are some papers on modeling probability density functions. These work typically apply some transformations to density functions and then employ linear models to the transformed functions. Brumback and Lindstrom [4] proposed a self-modeling method for monotone functions involving the transformation proposed by Jupp [17], which is a bijective map from the space of monotone increasing vectors to Euclidean space. Gervini [8] used the Jupp transformation to study warped functional regression. Peterson and Müller [22] proposed to use the log quantile density transformation and log hazard transformation to map a density function into a linear Hilbert space. Kokoszka et al. [20] used the same transformations to predict density functions. Another way is to study the manifold structure of warping functions. Here some of these related methods are reviewed. Cheng and Wu [6] used local linear regression models to study the scalar-on-manifold regression problem, where covariates lie on an unknown manifold. Su et al. [27] employed the transported square-root vector field (TSRVF) to implement statistical analysis of trajectories on Riemannian manifolds. Dai and Müller [7] developed principal component analysis for sphere-valued functional data. They proposed to apply functional principal component analysis (fPCA) to the tangent vectors at the Fréchet mean of the sphere.

However, all these methods have some limitations. One common characteristic of the first kind of method is that the transformations all involve the "logarithm" which sometimes dictates the need of another re-scaling step (e.g., $\log(f(Q(t)))$ and $\log(f(t)/\{1-F(t)\})$, where $F(t)$ and $Q(t)$ are the cumulative distribution function and quantile function of the density function $f(t)$, see Peterson and Müller [22]). A major limitation of the logarithm function is that it either shrinks the variation of large values or exaggerates the variation of values close to 0. In addition, density functions (and warping functions) lie in a nonlinear space, and it is always unnatural to use linear models directly. Regarding the second framework, one may consider applying linear models to the tangent space of the manifold composed of the square root of slope functions (SRSF) of warping functions $\gamma(t)$, defined as $\sqrt{\dot{\gamma}(t)}$. The SRSFs of warping functions lie on an infinite dimensional sphere, and thus the tangent space has clear and simple representation. However, the SRSFs of warping functions form only the positive orthant of the sphere ($-\sqrt{\dot{\gamma}(t)}$ is not included), and the predicted SRSFs by linear models may lie on the negative orthant. In addition, this approach still seeks to transform nonlinear space to linear space, and thus also changes the original variation. All of these problems motivate us to develop a new methodology to predict the stochastic process composed of warping functions.

We develop a novel method that can jointly predict amplitude and warping functions. The major advantage of our method is that it does not require any unnatural transformations and it retains the predicted warping functions strictly in their original non-linear space. To be more specific, we develop a state-space model where warping functions are assumed to be driven by hidden states, and consequently, there is no need to transform warping functions between linear and non-linear spaces.

We first implement functional registration to obtain amplitude and warping functions. To predict warping functions, we propose a state-space model, in which the states are driven by a Markov chain. Spherical K-means clustering, which is a popular technique for dimension reduction of non-linear space, is used to reveal the potential low dimensionality of warping functions. In the model, finite prototypes are employed to represent the nonlinear space of warping functions, where each warping function is assumed to be the sum of its corresponding prototype and a random error function. For the prediction of the amplitude function, we develop a varying coefficient operator functional auto-regressive model. Varying-coefficient models have been extended to functional data. Sentürk and Müller [24] generalized functional varying coefficient model to incorporate the influence of recent past values of predictors on current response. Further improvements were reported in Sentürk and Müller [25] which proposed a new representation for varying coefficient functions and introduced a smooth history index function to model the dependence of the response on the recent past values of predictors. Krafty et al. [21] employed a varying coefficient model in the analysis of tumor growth curves. Our varying coefficient model is fully functional, and the states of previous warping functions influence the current coefficient operators. The predicted warping functions and amplitude functions are combined to obtain the final prediction.

In this article, the following issues will be addressed:

1. Since the real states in the state-space model are unknown in practice, the transition probability matrix of the hidden Markov chain has to be estimated through the estimated states instead of the real states. The large-sample behavior of the estimator will be investigated in this paper.
2. A method for determining the dimension and order of the varying coefficient operator functional auto-regressive model will be developed.
3. We will develop a measure to evaluate the performance of our proposed method in preserving the common pattern.

The rest of the paper is organized as follows. In Section 2, we formulate the model for the stochastic process of warping functions and amplitude functions, the joint prediction procedure of amplitude and phase variation, and discuss how to measure shape similarity. In Section 3, we derive the asymptotic properties of the least squares estimator of the transition matrix in the state-space model. Section 4 displays the results of the simulation study comparing the prediction performance of the SP method and some other competitor methods. In Section 5, we report the results of the analysis on the NMHC concentration. Section 6 concludes the article.

2. Models, algorithms, and shape similarity

2.1. Amplitude and phase variation

In this section, we formulate the models for amplitude and phase variation. Let $\{f_n(t)\colon n \in \mathbb{N}\}$ be an arbitrary stationary functional time series defined on a common probability space $(\Omega, \mathcal{A}, P)$, where the function index n is discrete and the time index t is continuous. Assume the following decomposition $f_n(t) = X_n \circ \gamma_n(t)$. In this decomposition, $X_n(t)$ is the amplitude function and $\gamma_n(t)$ is the warping function. The observations $\{f_n(t)\colon n \in \mathbb{N}\}$ are elements of the Hilbert space $H = L^2[0,1]$ equipped with the inner product $\langle f_1, f_2\rangle = \int_0^1 f_1(t)f_2(t)dt$, and $f_n(t) < \infty$ for any $t \in [0,1]$. The norm of each f_n satisfies $\|f_n\|_2 = \sqrt{\langle f_n, f_n\rangle} < \infty$. Define the mean function and covariance function of the amplitude functions as follows

$$\mu(t) = E\{X_n(t)\}, \qquad K(t,s) = \operatorname{cov}\{X_n(t), X_n(s)\}.$$

In practice, $\mu(t)$ and $K(t,s)$ are always unknown and need to be estimated from samples $\{X_1(t), \ldots, X_N(t)\}$ as follows:

$$\hat{\mu}(t) = \frac{1}{N}\sum_{n=1}^{N} X_n(t), \qquad \widehat{K}(t,s) = \frac{1}{N}\sum_{n=1}^{N}\{X_n(t) - \hat{\mu}(t)\}\{X_n(s) - \hat{\mu}(t)\}.$$

By Mercer's theorem, $K(t,s)$ and $\widehat{K}(t,s)$ admit the following decomposition

$$K(t,s) = \sum_{m=1}^{\infty} \lambda_m \nu_m(t)\nu_m(s), \qquad \widehat{K}(t,s) = \sum_{m=1}^{\infty} \hat{\lambda}_m \hat{\nu}_m(t)\hat{\nu}_m(s),$$

where $\langle \nu_{m_1}, \nu_{m_2}\rangle = 0$ $(m_1 \neq m_2)$ and $\|\nu_m\|_2 = 1$ $(m \geq 1)$. The warping functions $\gamma_n\colon H \to H$ have the following property: $\gamma_n(0) = 0$, $\gamma_n(1) = 1$, γ_n is invertible, both γ_n and γ_n^{-1} are continuous, and assume that the first order derivative exists and satisfies $\dot{\gamma}_n(t) < \infty$ for all $t \in [0,1]$. Let Γ denote the set of all such functions. The square root of slope function (SRSF) of $\gamma_n(t)$ is defined as

$$s_n(t) = S(\gamma_n(t)) = \sqrt{\dot{\gamma}_n(t)},$$

and a SRSF $s_n(t)$ can be transformed back into a warping function $\gamma_n(t)$ by applying $S^{-1}(\cdot)$ to it

$$\gamma_n(t) = S^{-1}(s_n(t)) = \int_0^t s_n^2(u)du, \qquad 0 < t < 1.$$

Evidently $\|s_n(t)\|_2 = 1$, thus $\{s_n(t)\colon n \in \mathbb{N}\}$ lie on an infinite-dimensional sphere. In practice, only $\{f_n(t)\colon n \geq 1\}$ are observed, and we propose to apply functional registration algorithm (see e.g., Ramsay and Silverman [28], Kneip & Ramsay [19], and Srivastava & Klassen [26]) to obtain $\{X_n(t)\colon n \geq 1\}$ and $\{\gamma_n(t)\colon n \geq 1\}$. In the following, it is assumed that both $\{X_n(t)\colon n \geq 1\}$ and $\{\gamma_n(t)\colon n \geq 1\}$ are already obtained.

2.2. Functional auto-regressive model for amplitude functions

One way to model amplitude functions is by a FAR model. A FAR(q) process is defined by the stochastic recursion

$$X_n(t) - \mu(t) = \sum_{j=1}^{q} \Phi_j(X_{n-j} - \mu)(t) + \epsilon_n(t),$$

where $\{\epsilon_n(t)\colon n \in \mathbb{N}\}$ are centered, independent and identically distributed innovations in $H = L^2[0,1]$ and $\Phi_j(\cdot)\colon H \to H$ is a bounded linear operator for $j = 1, \ldots, q$, and are defined so that the above recursive equation has a unique causal solution (see [10], pp. 236). Horváth and Kokoszka [10] has developed a sufficient condition for causality of FAR(1) process, and the result can be extended to FAR(q) process ($q > 1$) under the state-space form of FAR(q) process.

The FAR model is easy-to-implement for the prediction of functional time series. One approach of estimation is to first project functions onto a finite dimensional sub-space spanned by some functional basis, e.g., functional principal components (fPC), then multivariate techniques are applied without much loss of information (see Aue et al. [2]).

2.3. State-space model for warping functions

Phase variation, which pertains to the variation of locations of curve features, is captured by the warping functions $\{\gamma_n(t)\colon n \in \mathbb{N}\}$. Since $\{\gamma_n(t)\colon n \in \mathbb{N}\}$ are defined in an infinite dimensional non-linear manifold, linear methods are not appropriate for the prediction of $\{\gamma_n(t)\colon n \in \mathbb{N}\}$. Note that it is computationally intractable to predict warping functions in the infinite dimensional manifold Γ. Hence, we propse to employ non-linear dimensional reduction techniques, and develop the state-space model with the following assumptions.

- The process $\{\gamma_n(t)\colon n \in \mathbb{N}\}$ is driven by a Markov chain, which is irreducible, ergodic, and has finite states. Each state c_n is associated with a fixed prototype warping function $b_{c_n}(t)$. $\gamma_n(t)$ is expressed as the summation of its corresponding prototype and a random error function $u_n(t)$.
- The random error functions $\{u_n(t)\colon n \in \mathbb{N}\}$ are of mean zero, and given c_n, $u_n(t)$ is independent of c_m and $u_m(t)$, $m \neq n$, and are constrained such that the resulting function $\gamma_n(t)$ is still a warping function.

Suppose the Markov chain has g states, then each state c_n can be represented by a state-indicating row vector ω_n, which is g-dimensional satisfying $\omega_{n,c_n} = 1$ and $\omega_{n,i} = 0$, for $i \neq c_n$. Denote P as the transition probability matrix. The state-space model is specified as follows

$$E[\omega_n|\omega_1, \ldots, \omega_{n-1}] = E[\omega_n|\omega_{n-1}] = \omega_{n-1}P,$$
$$\gamma_n(t) = \sum_{j}^{g} \omega_{n,j} b_j(t) + u_n(t).$$

The second equation is an analogue of the fPC representation of functions in $L^2[0,1]$. The prototypes $\{b_j(t)\colon j=1,\ldots,g\}$ can be viewed as a series of basis functions of Γ, and the state-space model is a discrete approximation of the continuous evolution of warping functions. The number of prototypes depends on the variation of warping functions. Under high variation, a large number of prototypes are typically needed to approximate the dynamic process $\{\gamma_n(t)\colon n\in\mathbb{N}\}$ well by the state-space process.

Remark 2.1. *In this state-space model, warping functions are driven by hidden states and it is not needed to employ linear models to account for its variation, and consequently, there is no need to transform warping functions between linear and non-linear spaces.*

Remark 2.2. *One possible concern is identifiability, say, $X\circ\gamma_1$ and $(X\circ\gamma_0^{-1})\circ(\gamma_0\circ\gamma_1)$ are the same, where γ_0 and γ_1 are two arbitrary different warping functions. However, once the template is fixed (e.g., sample mean) for functional registration, the decomposition is identifiable. Specifically, if the template $X_0(t)$ is fixed, the warping function $\gamma_n = \arg\min_{\gamma\in\Gamma}\|f_n - X_0\circ\gamma\|$ is unique for arbitrary n, where $\|\cdot\|$ is some metric employed for functional registration, and in this paper, we propose to use Fisher-Rao metric.*

2.3.1. Estimation of the state-space model

Since the hidden states and transition probability matrix are unknown in practice, we need to first estimate b_j's, ω_n's, and then P. We apply spherical K-means clustering, which is a widely-accepted dimension reduction technique for non-linear space, to the SRSFs of warping functions, and use the cluster centroids as the estimators of the SRSFs of b_j's. The estimators of b_j's are obtained by applying $S^{-1}(\cdot)$ to the cluster centroids,

$$\hat{b}_j(t) = S^{-1}(\hat{p}_j(t)), \qquad j=1,\ldots,g,$$

where $\hat{p}_j(t)$ is the centroid of the j-th cluster of SRSFs. The classified categories of $\{s_n(t)\colon n\in\mathbb{N}\}$ are considered as the estimated states of $\{\gamma_n(t)\colon n\in\mathbb{N}\}$. More details are discussed below.

The standard spherical K-means clustering aims to minimize

$$D=\sum_{n=1}^{N}(1-\cos(s_n,p_{c_n}))=\sum_{n=1}^{N}(1-\langle s_n,p_{c_n}\rangle)$$

over all assignments of objects n to cluster $c_n\in\{1,\ldots,g\}$ and over all SRSF representations of prototype warping functions $p_1,\ldots,p_g$. A typical projection and minimization procedure is repeated to obtain $\hat{c}_n$'s and $\hat{p}_j$'s.

Let $\hat{\omega}_n$ denote a g-dimensional vector where only the $\hat{c}_n$-th element is 1 and the rest elements are zeros. Then P is estimated by the least squares method, where ω_n is replaced with $\hat{\omega}_n$, say,

$$\widehat{P}=\arg\min_{P}\sum_{n=2}^{N}\|\hat{\omega}_n-\hat{\omega}_{n-1}P\|_2^2.$$

The number of hidden states is unknown in practice, and we propose a cross-validation method in Section 2.4.2 to select g. We assume the selected g is correct, and will not distinguish between the selected g and the real number of states. Note that, using the R package **skmeans**, spherical K-means clustering algorithm can be implemented by the R function *skmeans* (see Hornik et al. [9]). The estimation procedure is summarized in Algorithm 1:

Algorithm 1 Estimation of the state-space model

Step 1 Obtain the SRSFs of warping functions, $s_n = S(\gamma_n)$.
Step 2 Fix the number of states g, apply spherical K-means clustering to $\{s_n : n \in \mathbb{N}\}$, and obtain the cluster centroids $\{\hat{p}_j : j = 1, \ldots, g\}$ and the classified categories $\{\hat{c}_n : n \in \mathbb{N}\}$.
Step 3 Apply $S^{-1}(\cdot)$ to $\{\hat{p}_j : j = 1, \ldots, g\}$ to obtain the estimated prototype warping functions, say, $\hat{b}_j = S^{-1}(\hat{p}_j),\ j = 1, \ldots, g$.

2.4. Joint prediction methodology

After separating amplitude and phase components, it is natural to consider how to predict the two components jointly, as they are not necessarily independent of each other. Because warping functions and amplitude functions are defined in two different spaces, it is necessary to find a common space for these two kinds of functions in the joint prediction. To be more specific, we assume the amplitude and warping functions are jointly driven by a Markov process. The joint modeling procedure is discussed below.

2.4.1. Prediction of warping function

We convert the stochastic process of warping functions into a Markov chain by applying spherical K-means clustering to their corresponding SRSFs, as has been discussed in Section 2.3. In order to incorporate the dependence between phase and amplitude variation, we assume the same kind of state-space model for amplitude functions, and apply K-means clustering to estimate the hidden states of amplitude functions. Similarly, the classified categories are treated as the estimated hidden states. Figure 2 shows the framework, where ω represents the true state and $\hat{\omega}$ represents the estimated state, and superscripts (a) and (f) refer to amplitude and phase component respectively.

The two categorical sequences are combined to obtain a new sequence, $\hat{\omega}_n = (\hat{\omega}_n^{(f)} \otimes \hat{\omega}_n^{(a)})$, where $\otimes$ signifies the Kronecker product. Then apply the least squares method to estimate the transition matrix P of this combined estimated Markov chain, where P is a $g\ell \times g\ell$ matrix, g is the number of states of phase variation and ℓ is the number of states of amplitude variation. The predicted state is $\hat{\hat{\omega}}_{N+1} = \hat{\omega}_N \widehat{P}$, and the predictor of $\omega_{N+1}^{(f)}$ is obtained as

$$\hat{\hat{\omega}}_{N+1}^{(f)} = \hat{\omega}_N \widehat{P} J,$$

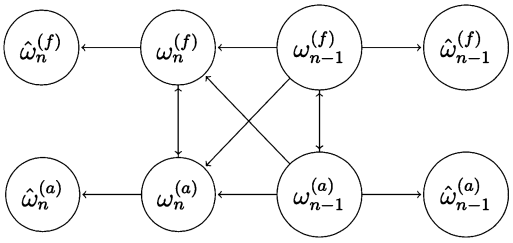

FIG 2. *Hidden states and estimated states.*

where J is the $(g\ell) \times g$ matrix

$$J = \begin{pmatrix} \mathbf{1}_\ell & \mathbf{0}_\ell & \cdots & \mathbf{0}_\ell \\ \mathbf{0}_\ell & \mathbf{1}_\ell & \cdots & \mathbf{0}_\ell \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_\ell & \mathbf{0}_\ell & \cdots & \mathbf{1}_\ell \end{pmatrix},$$

and $\mathbf{1}_\ell = (1, \ldots, 1)'_{1\times\ell}$, $\mathbf{0}_\ell = (0, \ldots, 0)'_{1\times\ell}$, and $'$ signifies transpose. The predicted warping function is $\hat{\gamma}_{N+1}(t) = \sum_{j=1}^{g} \hat{\hat{\omega}}_{N+1,j}^{(f)} \hat{b}_j(t)$. As a side note, $\hat{\hat{\omega}}_{N+1}$ is not a state-indicating vector, but a vector of probabilities of which the sum is one. Evidently $\hat{\gamma}_{N+1}(t)$ is a warping function since

$$\frac{d}{dt}\hat{\gamma}_{N+1}(t) = \sum_{j=1}^{g} \hat{\hat{\omega}}_{N+1,j}^{(f)} \frac{d\hat{b}_j(t)}{dt} \geq 0,$$

$$\sum_{j=1}^{g} \hat{\hat{\omega}}_{N+1,j}^{(f)} \hat{b}_j(1) = \sum_{j=1}^{g} \hat{\hat{\omega}}_{N+1,j}^{(f)} = 1, \qquad \sum_{j=1}^{g} \hat{\hat{\omega}}_{N+1,j}^{(f)} \hat{b}_j(0) = 0.$$

Remark 2.3. *When the sample size is small, some ad-hoc adjustments might be needed to let $\widehat{P}$ satisfy the constraints of a transition matrix. One approach is to obtain the transition matrix by solving the optimization problem*

$$\widehat{P} = \arg\min_{P \in P_{\mathcal{M}}} \|P - \widehat{P}^{\mathrm{LS}}\|_F,$$

where $P_{\mathcal{M}}$ is the set of all probability transition matrices, $\|\cdot\|_F$ is Frobenius norm, and $\widehat{P}^{\mathrm{LS}}$ is the original least squares estimator of P.

2.4.2. Data-driven selection of the number of states

To the best of our knowledge, there is no widely accepted procedure for order selection of hidden Markov models. The selection of state number is a trade-off between bias and variance. A large number of states decrease the approximation error by prototype warping functions, but increase variance of estimation.

Considering that our purpose is prediction, we propose an approach based on prediction error.

The prediction performance is evaluated by ℓ^2 mean squared error and amplitude distance (Section 2.5.2). Assume that there is a large test data-set D_{test} which is an independent copy of the dataset used for model fitting, and use the first 80% curves in D_{test} to fit a model with g phase states and ℓ amplitude states, and predict the rest 20% curves with the fitted model. Next calculate the mean squared error and the average amplitude distance between the predicted curves and the curves to be predicted, and refer to these two errors for the order selection. In practice, the sample size may be limited, and it is not possible to reserve a large fraction of data for the testing set. In this case, Monte-Carlo cross-validation is a good alternative approach. A fraction of consecutive curves are selected as training set and the rest curves are used for testing. This procedure is repeated multiple times where the partitions are randomly chosen on each run. A group of candidate state numbers are preset, and the two average errors are computed for models with different candidates. The state numbers are selected such that both errors are decent.

2.4.3. Prediction of amplitude function

We now develop the FAR model with varying coefficient operators for the prediction of amplitude functions. The coefficient operator is determined by the state of the previous warping function. Define $Y_n(t) = X_n(t) - \mu(t)$ and let $c_n^{(f)}$ be the hidden state of γ_n. The proposed model has the following representation

$$Y_{n+1}(t) = \sum_{h=1}^{q} \Phi_h^{(c_n^{(f)})}(Y_{n+1-h})(t) + \epsilon_{n+1}(t),$$

where $\{\epsilon_n(t)\colon n \in \mathbb{N}\}$ are centered, independent and indentically distributed innovations in $L^2[0,1]$, and $\{\Phi_h^{(k)}\colon k = 1,\dots,g, h = 1,\dots,q\}$ are bounded linear operators, and are constrained so that the above recursive equation has a unique causal solution.

The estimation of $\{\Phi_h^{(k)}\colon k = 1,\dots,g, h = 1,\dots,q\}$ is obtained by minimizing the objective function

$$S(\Phi) = \sum_{n=h}^{N-1} \left\| Y_{n+1} - \sum_{h=1}^{q} \Phi_h^{(c_n^{(f)})}(Y_{n+1-h}) \right\|_2^2.$$

By simple decomposition,

$$S(\Phi) = \sum_{k=1}^{g} \sum_{n_k=1}^{N_k} \left\| Y_{n_k+1} - \sum_{h=1}^{q} \Phi_h^{(k)}(Y_{n_k+1-h}) \right\|_2^2,$$

where N_k is the number of Y_{n+1} so that γ_n is in state k. Then minimize the

following quantity to obtain the estimation of $\{\Phi_h^{(k)}\}_{h=1}^p$:

$$S_k(\Phi) = \sum_{n_k=1}^{N_k} \left\| Y_{n_k+1} - \sum_{h=1}^{q} \Phi_h^{(k)}(Y_{n_k+1-h}) \right\|_2^2.$$

After projecting all functional elements onto the sub-eigenspace spanned by the finite major functional principal components of $\{X_n(t)\colon n \in \mathbb{N}_+\}$, the multivariate technique are applied to estimate $\{\Phi_h^{(k)}\colon h = 1, \ldots, q\}$ for each k. Denote $\widehat{\Phi}_h^{(k)}$ as the estimator of $\Phi_h^{(k)}$, then the predictor of Y_{N+1} is

$$\widehat{Y}_{N+1} = \sum_{k=1}^{g} \sum_{h=1}^{q} \widehat{\Phi}_h^{(k)}(Y_{N+1-h}) \mathbb{1}(\hat{c}_N = k).$$

The entire joint prediction procedure is summarized in Algorithm 2.

Algorithm 2 Joint prediction algorithm (one-step ahead)

Step 1 Apply functional registration algorithm to obtain the amplitude and warping functions.
Step 2 Apply spherical K-means clustering algorithm resp. K-means clustering algorithm to the SRSFs of the warping functions resp. the amplitude functions to obtain the estimated states. Combine the two state sequences, fit a Markov model, and obtain the prediction of the next warping function $\hat{\gamma}_{N+1}$ by the state space model.
Step 3 Obtain the prediction of the next amplitude function, $\widehat{Y}_{N+1}$, based on a FAR model with varying coefficient operators.
Step 4 Warp $\widehat{Y}_{N+1} + \hat{\mu}$ by $\hat{\gamma}_{N+1}$ to obtain the final prediction, $\hat{f}_{N+1} = (\widehat{Y}_{N+1} + \hat{\mu}) \circ \hat{\gamma}_{N+1}$.

Remark 2.4. *The final expression is binary. In practice, the following weighted predictor can also be considered,*

$$\widehat{Y}_{N+1} = \sum_{k=1}^{g} \sum_{h=1}^{q} \widehat{\Phi}_h^{(k)}(Y_{N+1-h}) P(\hat{c}_N = k).$$

The weighted predictor has smaller variance but larger bias. The probabilities of states $P(\hat{c}_N = k)$ need to be estimated under some model, for example, $P(\hat{c}_N = k) \propto 1/d(\hat{\gamma}_N, \hat{b}_k)$, where $d(\hat{\gamma}_N, \hat{b}_k)$ is some distance between $\hat{\gamma}_N$ and $\hat{b}_k$.

2.4.4. Parameter selection

Now we develop the functional final prediction error (fFPE) criterion to select the order and dimension of the sub-eigenspace for the prediction of amplitude functions. Create the d-variate fPC score vector $\boldsymbol{Y}_n = (y_{n,1}, \ldots, y_{n,d})'$, where $y_{n,m} = \langle Y_n, \nu_m \rangle = \langle X_n - \mu, \nu_m \rangle$. Since the eigenfunctions are orthogonal and the fPC scores are uncorrelated for each Y_n, the mean squared prediction error is decomposed as

$$E\left[\left\| Y_{N+1} - \widehat{Y}_{N+1} \right\|^2 \right] = E\left[\left\| \sum_{m=1}^{\infty} y_{N+1,m} \nu_m - \sum_{m=1}^{d} \hat{y}_{N+1,m} \nu_m \right\|^2 \right]$$

$$= E\left[\left\|\boldsymbol{Y}_{N+1} - \widehat{\boldsymbol{Y}}_{N+1}\right\|^2\right] + \sum_{m>d} \lambda_m,$$

where $\|\cdot\|$ denotes the ℓ^2-norm, and $\hat{y}_{N+1,m}$ is the prediction of $y_{N+1,m}$ from the past d-variate fPC score vectors. As for the first summand, assume $\{\boldsymbol{Y}_n : n \in \mathbb{N}\}$ follows a d-variate VAR(q) process (see Aue et al. [2] for the justification of the VAR process) with varying coefficient matrix, that is, $\boldsymbol{Y}_{n+1} = \boldsymbol{\Phi}_1^{(c_n^{(f)})}\boldsymbol{Y}_n + \ldots + \boldsymbol{\Phi}_q^{(c_n^{(f)})}\boldsymbol{Y}_{n-q+1} + \boldsymbol{Z}_{n+1}$, where $\boldsymbol{Z}_n$ is the error term. For any state of warping function k, it can be shown that (see, e.g., Lütkepohl [14]) $\sqrt{N_k}(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) \xrightarrow{\mathcal{L}} \mathcal{N}_{qd^2}(0, \Sigma_{Z,k}^d \otimes \Gamma_{q,k}^{-1})$, where $\boldsymbol{\beta}_k = \mathrm{vec}([\boldsymbol{\Phi}_1^{(k)}, \ldots, \boldsymbol{\Phi}_q^{(k)}]')$ and $\hat{\boldsymbol{\beta}}_k$ is the least squares estimator of $\boldsymbol{\beta}_k$, $\Sigma_{Z,k}^d$ is the covariance matrix of $\boldsymbol{Z}_{n+1}$ as $c_n^{(f)} = k$, $\Gamma_{q,k} = \mathrm{var}(\mathrm{vec}([\boldsymbol{Y}_n, \ldots, \boldsymbol{Y}_{n-q+1}]))$ as $c_n^{(f)} = k$, and $\xrightarrow{\mathcal{L}}$ signifies convergence in distribution. Let $\widehat{\boldsymbol{Y}}_{N+1}^{(k)}$ be the predictor of $\boldsymbol{Y}_{N+1}$ as $c_n^{(f)} = k$. Assuming the classification is correct, it follows that

$$\begin{aligned}
&E\left[\left\|\boldsymbol{Y}_{N+1} - \widehat{\boldsymbol{Y}}_{N+1}\right\|^2\right] = E\left[\left\|\boldsymbol{Y}_{N+1} - \sum_{k=1}^{g} \widehat{\boldsymbol{Y}}_{N+1}^{(k)} \mathbb{1}(c_N^{(f)} = k)\right\|^2\right] \\
&= E\left[E\left[\left\|\boldsymbol{Y}_{N+1} - \widehat{\boldsymbol{Y}}_{N+1}^{(c_N^{(f)})}\right\|^2 \middle| c_N^{(f)}\right]\right] = \sum_{k=1}^{g} E\left[\left\|\boldsymbol{Y}_{N+1} - \widehat{\boldsymbol{Y}}_{N+1}^{(k)}\right\|^2\right] P(c_N^{(f)} = k) \\
&= \sum_{k=1}^{g} E\left[\left\|\boldsymbol{Y}_{N+1} - \sum_{h=1}^{q} \widehat{\boldsymbol{\Phi}}_h^{(k)} \boldsymbol{Y}_{N+1-h}\right\|^2\right] P(c_N^{(f)} = k) \\
&= \sum_{k=1}^{g} \left\{\mathrm{tr}(\Sigma_{Z,k}^d) + E\left[\left\|\sum_{h=1}^{q} (\boldsymbol{\Phi}_h^{(k)} - \widehat{\boldsymbol{\Phi}}_h^{(k)}) \boldsymbol{Y}_{N+1-h}\right\|^2\right]\right\} P(c_N^{(f)} = k) \\
&= \sum_{k=1}^{g} \left\{\mathrm{tr}(\Sigma_{Z,k}^d) + E\left[\left\|I_p \otimes (\boldsymbol{Y}_N', \ldots, \boldsymbol{Y}_{N-q+1}')(\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k)\right\|^2\right]\right\} P(c_N^{(f)} = k) \\
&\sim \sum_{k=1}^{g} \left\{\mathrm{tr}(\Sigma_{Z,k}^d) + \frac{qd}{N_k}\mathrm{tr}(\Sigma_{Z,k}^d)\right\} P(c_N^{(f)} = k),
\end{aligned}$$

where $a_N \sim b_N$ means $a_N/b_N \to 1$. Finally we conclude that

$$E[\|Y_{N+1} - \widehat{Y}_{N+1}\|^2] \sim \sum_{k=1}^{g} \left(\frac{N_k + qd}{N_k}\right) \mathrm{tr}(\Sigma_{Z,k}^d) P(c_N^{(f)} = k) + \sum_{m>d} \lambda_m.$$

Replacing $\mathrm{tr}(\Sigma_{Z,k}^d)$ with $\mathrm{tr}(\widehat{\Sigma}_{Z,k}^d)$, $P(c_N^{(f)} = k)$ with N_k/N, and λ_m with $\hat{\lambda}_m$, where $\widehat{\Sigma}_{Z,k}^d$ is the unbiased estimator of $\Sigma_{Z,k}^d$, the fFPE criterion is given by,

$$\mathrm{fFPE}(q,d) = \sum_{k=1}^{g} \left(\frac{N_k + qd}{N}\right) \mathrm{tr}(\widehat{\Sigma}_{Z,k}^d) + \sum_{m>d} \hat{\lambda}_m.$$

We propose to select q and d by minimizing $\mathrm{fFPE}(q,d)$.

2.5. Shape similarity

2.5.1. Functional shape space

One of the main questions considered in this article is: what is a good measurement of shape similarity? In order to compare the shapes of different trajectories, we need to formally define the functional shape space $\mathcal{E}$ and to evaluate shape similarity. Here, we shall follow the convention that shape is independent of scale and location. We first rescale and relocate functions, so that they are of unit norm, and start at the same value. Then we study the shape difference of the thus obtained set. This resulting space is termed pre-shape space.

Suppose there are two functions f_1 and f_2, with the corresponding transformations in the pre-shape space as $\tilde{f}_1$ and $\tilde{f}_2$. We propose the principle that, if $\tilde{f}_1$ can be warped into $\tilde{f}_2$, the two functions f_1 and f_2 are considered to be of the same shape. This idea is motivated by shape data analysis (see e.g., Srivastava and Klassen [26]). To be specific, stretching, rotating, or relocating do not change the shape of planar shape objects. As a motivating example in shape data analysis, suppose that there is a planar contour delineating a human hand, stretching, rotating, and relocating the contour will not change the shape of human hand.

In the functional shape space, we unify the shape representations, that is, obtain the unification of all points in pre-shape space representing the same shape. Therefore, the functional shape space $\mathcal{E}$ is defined as the quotient space of $L^2[0,1]$ with respect to relocating, rescaling and warping. We define the equivalence relation $\equiv$ on $\mathcal{E}$ as follows: let $\tilde{f}_1$, $\tilde{f}_2$ be the pre-shape elements of two functions f_1, f_2, then $f_1 \equiv f_2$ if there exists a warping function γ such that $\tilde{f}_1 = \tilde{f}_2 \circ \gamma$. For any function f_0, the set of all functions, of which transformations in the pre-shape space can be warped into $\tilde{f}_0$, is considered as an object in the functional shape space $\mathcal{E}$, that is, $[f_0] = \{f\colon \tilde{f} \circ \gamma = \tilde{f}_0, \gamma \in \Gamma\} \in \mathcal{E}$. Based on this definition, the distance $d([f_1],[f_2])$ between two shape objects $[f_1]$ and $[f_2]$ should be invariant to relocating, rescaling and warping of f_1 and f_2.

2.5.2. Amplitude distance

For any $f \in H_0 = \{f \in H\colon \dot{f} > 0\}$, and $\nu_1, \nu_2 \in T_f(H)$, where $T_f(H)$ is the tangent space of H at f, defined as $\{h \in H\colon \langle f, h\rangle = 0\}$, the Fisher–Rao metric is defined as the inner product

$$\langle\langle \nu_1, \nu_2 \rangle\rangle_f = \frac{1}{4}\int_0^1 \dot{\nu}_1(t)\dot{\nu}_2(t)\frac{1}{\dot{f}(t)}dt.$$

One important property of Fisher-Rao metric is invariance of simultaneous warping: for any $\gamma \in \Gamma$, $d_{\mathrm{FR}}(f_1, f_2) = d_{\mathrm{FR}}(f_1 \circ \gamma, f_2 \circ \gamma)$, where d_{FR} denotes the geodesic distance induced by the Fisher-Rao metric. Under the SRSF representation, the Fisher–Rao Riemannian metric on H_0 becomes the standard ℓ^2-metric (see [26], pp. 106). With this property, the geodesic distance under

the Fisher–Rao metric can be written explicitly as $d_{\mathrm{FR}}(f_1, f_2) = \|s_1 - s_2\|_2$, where s_1, s_2 are the SRSF representations of f_1, f_2. The Fisher–Rao metric is defined only on a subset $H_0 \subset H$, but under SRSF representation, it can be generalized to H endowed with the ℓ^2-metric. The ℓ^2-metric on SRSF representation space is termed as the extended Fisher–Rao metric.

We shall use the amplitude distance (2.1), which has been shown to be a proper distance in the functional shape space, to measure shape similarity,

$$d([f_1], [f_2]) = \inf_{\gamma} d_{\mathrm{FR}}(\tilde{f}_1, \tilde{f}_2 \circ \gamma). \tag{2.1}$$

If two functions are of the same shape, then the amplitude distance between the two functions is zero. The geodesic distance induced by the Fisher–Rao metric is invariant to simultaneous warpings. Therefore, the effect of phase variation does not influence the amplitude distance between two functions, say, for any two different warping functions γ_1 and γ_2, $\inf_\gamma d_{\mathrm{FR}}(\tilde{f}_1 \circ \gamma_1, \tilde{f}_2 \circ \gamma_2 \circ \gamma) = \inf_\gamma d_{\mathrm{FR}}(\tilde{f}_1, \tilde{f}_2 \circ \gamma)$, and thus the amplitude distance between two shape objects is unique (see [26], pp. 85–88).

Remark 2.5. *In this paper, we use both the amplitude distance and the Euclidean distance to evaluate the prediction.*

3. Theoretical results

The least squares method is employed to estimate the unknown transition probabilities, and aim to find the asymptotic properties of the estimator. It is known that the least squares estimator of the transition matrix of a Markov chain is consistent and asymptotically normal (see van der Plas [29]). However, since the real hidden states need to be estimated, the least squares estimator of the transition matrix P is not necessarily consistent with P. To find the matrix that $\widehat{P}$ is consistent with, the following assumptions are needed.

A1. The Markov chain $\{\omega_n \colon n \in \mathbb{N}\}$ is stationary and ergodic, and has finite states;

A2. The estimated prototypes are obtained from an independent copy of observations, and the estimated state $\hat{\omega}_n^{(a)}$ resp. $\hat{\omega}_n^{(f)}$ is independent of $\mathcal{F}_{a,0}^{\infty}$ and $\mathcal{F}_{f,0}^{\infty}$ given $\omega_n^{(a)}$ resp. $\omega_n^{(f)}$, where $\mathcal{F}_{a,0}^{\infty} = \sigma(\omega_0^{(a)}, \hat{\omega}_0^{(a)}, \ldots, \omega_\infty^{(a)}, \hat{\omega}_\infty^{(a)})$ and $\mathcal{F}_{f,0}^{\infty} = \sigma(\omega_0^{(f)}, \hat{\omega}_0^{(f)}, \ldots, \omega_\infty^{(f)}, \hat{\omega}_\infty^{(f)})$, and $\sigma(X)$ signifies the σ-algebra induced by X;

A3. $p(\hat{c}_n^{(f)} = \beta | c_n^{(f)} = \alpha)$ are the same across n for any $\alpha, \beta = 1, \ldots, g$, and $p(\hat{c}_n^{(a)} = \beta' | c_n^{(a)} = \alpha')$ are the same across n for any $\alpha', \beta' = 1, \ldots, \ell$.

Note that Assumption (A2) is compatible with the assumption on the error term u_n of the state-space model. Based on the model assumption, the estimated state $\hat{\omega}_n$ is only relevant to the real state ω_n and the random error u_n, so Assumption (A2) is a natural consequence of the assumption on u_n. Assumption

(A2) means, given the corresponding real state, the estimated state is independent of all other states. This is a reasonable assumption, since as the sample size grows large enough, the estimated prototype functions tend to be uncorrelated with any individual function. Assumption (A3) guarantees a constant transition probability matrix of the estimate states.

The Bayesian theorem implies the following proposition.

Proposition 3.1. *Under Assumptions (A1)–(A3), the transition probabilities of the combined estimated process* $\{\hat{\omega}_n^{(f)} \otimes \hat{\omega}_n^{(a)} : n \in \mathbb{N}\}$ *are given by*

$$P(\hat{\omega}_{n+1}^{(f)}, \hat{\omega}_{n+1}^{(a)} | \hat{\omega}_n^{(f)}, \hat{\omega}_n^{(a)}) = \sum_{\omega_{n+1}^{(f)}, \omega_{n+1}^{(a)}, \omega_n^{(f)}, \omega_n^{(a)}} P(\omega_{n+1}^{(f)}, \omega_{n+1}^{(a)} | \omega_n^{(a)}, \omega_n^{(f)}) P(\hat{\omega}_{n+1}^{(a)} | \omega_{n+1}^{(a)}) P(\hat{\omega}_{n+1}^{(f)} | \omega_{n+1}^{(f)}) \times \frac{P(\hat{\omega}_n^{(a)} | \omega_n^{(a)}) P(\hat{\omega}_n^{(f)} | \omega_n^{(f)}) P(\omega_n^{(a)}, \omega_n^{(f)})}{\sum_{\omega_n^{(a)}, \omega_n^{(f)}} P(\hat{\omega}_n^{(a)} | \omega_n^{(a)}) P(\hat{\omega}_n^{(f)} | \omega_n^{(f)}) P(\omega_n^{(f)}, \omega_n^{(a)})}.$$

Remark 3.1. *Proposition 3.1 implies the transition probability of the estimated Markov chain.*

Next we show that the least squares estimator $\widehat{P}$ is consistent with $\widetilde{P} = \{P(\hat{\omega}_{n+1}^{(f)}, \hat{\omega}_{n+1}^{(a)} | \hat{\omega}_n^{(f)}, \hat{\omega}_n^{(a)})\}$. Notationally, let $L_N(P) = N^{-1} \sum_{n=2}^{N} \|\hat{\omega}_n - \hat{\omega}_{n-1} P\|_2^2$, then we develop the following theorem for the least squares estimator $\widehat{P}_N$, which is a generalization of the result of van der Plas [29].

Theorem 3.1. *Under Assumptions (A1)–(A3), for each N there exists a random matrix* $\widehat{P}_N$ *such that* $L_N(\widehat{P}_N) = \inf_P L_N(P)$ *and* $\lim_{N \to \infty} \widehat{P}_N = \widetilde{P}$ *a.s.*.

In order to establish the asymptotic normality of the least squares estimator $\widehat{P}_N$, we make one additional assumption as follows,

A4. The matrix $A = \{a_{ij}\}$ where $a_{ij} = 2E\{\langle \frac{\partial \hat{\omega}_0 P}{\partial \theta_i}|_{\widetilde{P}}, \frac{\partial \hat{\omega}_0 P}{\partial \theta_j}|_{\widetilde{P}} \rangle\}$ is positive definite, where θ_i is the i-th element of vec(P).

and introduce the following notations,

$$F_i(n, \theta) = \left\langle \hat{\omega}_n - \hat{\omega}_{n-1} P, \hat{\omega}_{n-1} \frac{\partial P}{\partial \theta_i} \right\rangle,$$
$$F_i(n, \tilde{\theta}) = \left\langle \hat{\omega}_n - \hat{\omega}_{n-1} \widetilde{P}, \hat{\omega}_{n-1} \frac{\partial P}{\partial \theta_i}\Big|_{\widetilde{P}} \right\rangle.$$

The asymptotic normality of the least squares estimator $\widehat{P}_N$ is established in Theorem 3.2.

Theorem 3.2. *Under Assumptions (A1)–(A4),*

$$N^{1/2}(\hat{\theta}_N - \tilde{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, A^{-1} \Sigma A^{-1}),$$

where $\tilde{\theta} = \mathrm{vec}(\widetilde{\mathrm{P}})$, $\hat{\theta}_N = \mathrm{vec}(\widehat{\mathrm{P}}_{\mathrm{N}})$, *and*

$$\Sigma_{ij} = E\{F_i(0,\tilde{\theta})F_j(0,\tilde{\theta})\} + 2\sum_{k=1}^{\infty} E\{F_i(0,\tilde{\theta})F_j(k,\tilde{\theta})\}.$$

Remark 3.2. *The estimation of the transition probability matrix is consistent and asymptotically normal. Therefore, the estimation behaves stably under large sample size.*

4. Simulations

Finite sample simulations were implemented to illustrate the effectiveness of the SP method. The method was tested on a FAR(1) process with phase variation. In each simulation run, 300 (or 600) functions were simulated, and the first 90% of the simulated trajectories were used for model fitting to do one-step ahead prediction for the remaining 10% of trajectories by a moving-window approach. Each simulation run was repeated 200 times. We compared our method with the prediction method of Aue et al. [2], which does not incorporate functional registration, through two kinds of distance, the ℓ^2 Euclidean distance (ℓ^2) and the amplitude distance (FR). We also compared our proposed state-space model with the transformation methods on the prediction of warping functions.

4.1. First simulation setup

4.1.1. Simulation of warping function

Based on the properties of B-splines (de Boor [5]), we develop the following procedure to simulate warping functions. We first generated four prototype warping functions with 7 B-splines. The B-spline scores of the four prototypes were generated through the following procedure:

1. Four 6-variate vectors with positive elements, $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{i6})$, $i = 1, 2, 3, 4$, were specified as follows:

$$\boldsymbol{\xi}_1 = (1.0, 1.2, 1.4, 1.6, 1.8, 2.0), \qquad \boldsymbol{\xi}_2 = (2.0, 1.8, 1.6, 1.4, 1.2, 1.0),$$
$$\boldsymbol{\xi}_3 = (0.3, 0.3, 1.2, 1.2, 0.3, 0.3), \qquad \boldsymbol{\xi}_4 = (1.2, 1.2, 0.3, 0.3, 1.2, 1.2).$$

2. The vectors obtained in the first step were transformed as follows:

$$\phi_{i,j+1} = \frac{\sum_{k=1}^{j} \xi_{ik}}{\sum_{k=1}^{6} \xi_{ik}}, \qquad j = 1, 2, \ldots, 6,$$

then concatenate a zero to each of the vectors $(\phi_{i2}, \ldots, \phi_{i7})$ for $i = 1, 2, 3, 4$ to finalize the score vectors of the prototype warping functions.

The four score vectors $\boldsymbol{\phi}_i = (\phi_{i1}, \ldots, \phi_{i7})$ satisfy $\phi_{i1} = 0$, $\phi_{i7} = 1$ and $\phi_{i1} < \phi_{i2} < \ldots < \phi_{i7}$. The prototypes were generated with 7 B-splines

$$b_i(t) = \sum_{j=1}^{7} \phi_{ij} B_j(t), \ t \in [0, 1].$$

The independent error warping functions, denoted by γ_n^e, were simulated through the same procedure, say, first simulate a 6-dimensional vector $\boldsymbol{\xi}_n^e$, then transform it to $\boldsymbol{\phi}_n^e$ and take the B-spline expansion. The elements in $\boldsymbol{\xi}_n^e$ independently follow the uniform distribution $U[1, 2]$. The state of warping functions were simulated under a Markov process with four states, and the probability transition matrix has the representation

$$P = \begin{pmatrix} p & (1-p)/3 & (1-p)/3 & (1-p)/3 \\ (1-p)/3 & p & (1-p)/3 & (1-p)/3 \\ (1-p)/3 & (1-p)/3 & p & (1-p)/3 \\ (1-p)/3 & (1-p)/3 & (1-p)/3 & p \end{pmatrix}.$$

Each state is associated with a prototype. The final warping functions were obtained by

$$\gamma_n(t) = (1-\tau) b_{c_n^{(f)}}(t) + \tau \gamma_n^e(t),$$

where $0 < \tau < 1$ determining the proportion of signal, and $c_n^{(f)}$ is the simulated state of the n-th warping function. Figure 3 displays the simulated warping functions and the prototypes.

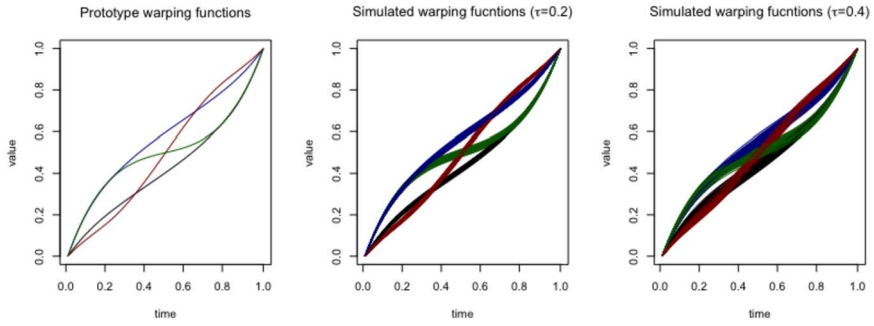

FIG 3. *Prototypes and simulated warping functions for different τ's*

4.1.2. Simulation of amplitude function

Amplitude functions were simulated with the same 7 B-splines, where the scores of the third and the fifth B-splines are significantly larger than those of the other B-splines. Thus all curves have the same two-peak pattern. The

two pronounced scores jointly follow a VAR(1) process with varying coefficient matrix, and the amplitude functions were obtained by the basis expansion $a_n(t) = \sum_{j=1}^{7} \zeta_{nj} B_j(t)$. The VAR(1) process has 4 coefficient matrices specified below

$$\begin{pmatrix} 0.8 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0.8 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0.8 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0.8 \end{pmatrix},$$

and the coefficient matrices were determined by the state of warping function. The varying coefficient VAR(1) model is specified below,

$$\begin{pmatrix} \zeta_{n+1,3} - 4 \\ \zeta_{n+1,5} - 6 \end{pmatrix} = \Phi^{(c_n^{(f)})} \begin{pmatrix} \zeta_{n,3} - 4 \\ \zeta_{n,5} - 6 \end{pmatrix} + e_{n+1},$$

where $e_n \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma)$, $\Sigma = 0.2\boldsymbol{I}_2$, and $\boldsymbol{I}_2$ is a 2×2 identity matrix. The other scores independently follow $\mathcal{N}(1, 0.1)$. The functional time series trajectories were obtained by applying the warping functions to the amplitude functions, $f_n(t) = a_n \circ \gamma_n(t)$.

Figure 4 displays the simulated amplitude functions and the simulated functional time series for different τ's. Table 1 and 2 display the average ℓ^2 prediction error (ℓ^2, defined as $\sum_n \|f_n - \widehat{f}_n\|_2/(0.1N)$) and amplitude difference (FR) between the predicted functions and the corresponding actual functions being predicted for $p = 0.5, 0.7, 0.9$ and $N = 300, 600$.

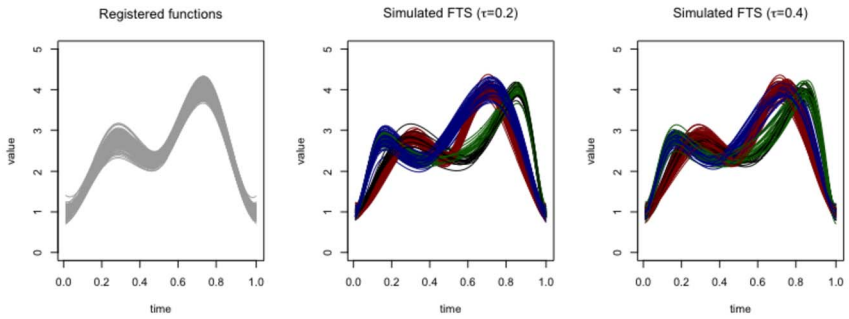

FIG 4. *Simulated curves for different τ's*

4.2. Second simulation setup

In the second setup, the amplitude functions were simulated similarly with one coefficient matrix $\Phi = 0.8\boldsymbol{I}_2$. The major difference is the simulation of warping functions. In this simulation setup, the same procedure is applied to simulate a sequence of independent warping functions $\{\gamma_n^e(t) \colon n \in \mathbb{N}\}$, where $\xi_{n,j}^e \overset{i.i.d}{\sim} U[0.5, 3]$ for $j = 1, \ldots, 6$, and then take the moving average of these functions to obtain $\{\gamma_n(t) \colon n \in \mathbb{N}\}$, say, $\gamma_n = \beta\gamma_{n-1}^e + (1-\beta)\gamma_n^e$, where β

TABLE 1
Average amplitude distance and ℓ^2 prediction error ($\tau = 0.4$). The format of each block: average (standard deviation $\times 100$)

p	0.5		0.7		0.9	
Shape-preserving prediction						
$N = 300$						
g	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
3	0.298(1.86)	0.192(1.08)	0.257(2.37)	0.199(1.12)	0.202(3.69)	0.196(1.07)
4	0.294(1.75)	0.193(1.05)	0.241(2.46)	0.199(1.02)	0.168(2.01)	0.196(0.94)
5	0.297(1.79)	0.193(0.95)	0.245(2.36)	0.200(0.95)	0.170(2.29)	0.196(1.13)
$N = 600$						
g	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
3	0.300(1.26)	0.189(0.75)	0.260(1.73)	0.198(0.73)	0.200(2.50)	0.194(0.85)
4	0.293(1.24)	0.190(0.74)	0.242(1.69)	0.200(0.73)	0.166(1.51)	0.193(0.69)
5	0.295(1.31)	0.191(0.71)	0.240(1.78)	0.199(0.71)	0.166(1.40)	0.193(0.78)
Prediction without registration						
$N = 300$						
–	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
–	0.289(1.85)	0.235(1.15)	0.245(2.12)	0.216(1.07)	0.185(2.23)	0.201(1.13)
$N = 600$						
–	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
–	0.290(1.27)	0.236(0.83)	0.247(1.61)	0.214(0.80)	0.184(1.35)	0.199(0.80)

TABLE 2
Average amplitude distance and ℓ^2 prediction error ($\tau = 0.2$). The format of each block: average (standard deviation $\times 100$)

p	0.5		0.7		0.9	
Shape-preserving prediction						
$N = 300$						
g	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
3	0.377(2.18)	0.205(1.27)	0.318(3.03)	0.202(1.22)	0.219(5.25)	0.206(1.62)
4	0.371(2.25)	0.203(1.12)	0.291(3.14)	0.201(1.29)	0.166(2.40)	0.207(1.71)
5	0.371(2.25)	0.205(1.10)	0.292(3.41)	0.203(1.26)	0.168(2.46)	0.207(1.73)
$N = 600$						
g	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
3	0.381(1.39)	0.206(0.81)	0.313(2.42)	0.202(0.99)	0.221(4.06)	0.209(1.24)
4	0.372(1.45)	0.203(0.88)	0.289(2.14)	0.200(0.89)	0.169(1.92)	0.208(1.41)
5	0.373(1.65)	0.203(0.90)	0.291(2.45)	0.201(0.86)	0.168(1.89)	0.207(1.40)
Prediction without registration						
$N = 300$						
–	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
–	0.352(2.35)	0.317(1.42)	0.285(2.71)	0.260(1.58)	0.179(2.77)	0.202(1.67)
$N = 600$						
–	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
–	0.354(1.45)	0.318(1.15)	0.279(2.31)	0.256(1.07)	0.176(1.95)	0.198(1.19)

takes value in $(0.3, 0.5, 0.7)$. Here, the amplitude and warping functions were predicted separately. Table 3 shows the average ℓ^2 prediction error and amplitude distance between the predicted curves and the corresponding real curves for different values of β and N.

TABLE 3
Average amplitude distance and ℓ^2 prediction error for different β. The format of each block: average (standard deviation $\times 100$)

β	0.3		0.5		0.7	
Shape-preserving prediction						
$N = 300$						
g	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
4	0.309(2.89)	0.202(1.11)	0.280(2.55)	0.198(1.15)	0.310(2.88)	0.202(1.21)
6	0.309(2.67)	0.203(1.10)	0.282(2.66)	0.199(1.09)	0.309(2.90)	0.204(1.11)
8	0.308(2.88)	0.205(1.05)	0.278(2.66)	0.201(1.16)	0.310(2.60)	0.204(1.07)
10	0.311(2.96)	0.205(1.03)	0.279(2.60)	0.201(1.11)	0.310(2.96)	0.206(1.15)
$N = 600$						
g	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
4	0.310(1.99)	0.201(0.81)	0.279(1.92)	0.197(0.72)	0.306(2.04)	0.200(0.82)
6	0.308(2.01)	0.202(0.77)	0.276(1.88)	0.197(0.71)	0.307(1.87)	0.202(0.80)
8	0.304(2.03)	0.202(0.85)	0.279(1.85)	0.198(0.78)	0.306(1.95)	0.203(0.84)
10	0.305(1.88)	0.202(0.84)	0.277(1.91)	0.197(0.69)	0.306(1.85)	0.203(0.79)
Prediction without registration						
$N = 300$						
–	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
–	0.301(2.76)	0.262(1.60)	0.267(2.41)	0.234(1.49)	0.301(2.65)	0.261(1.64)
$N = 600$						
–	ℓ^2	FR	ℓ^2	FR	ℓ^2	FR
–	0.301(1.87)	0.261(1.12)	0.265(1.78)	0.232(0.98)	0.297(1.70)	0.260(1.09)

4.3. Discussion on the simulations

In the first simulation setting, the optimal number of hidden states of warping functions is 4. Therefore, as g changes from 3 to 4, the performance of the SP method is significantly improved. Tables 1, 2 and 3 show that (1.) The SP method preserves the common pattern after incorporating functional registration into prediction, and (2.) The performance of the SP method is robust to the selection of the number of hidden states. When the phase variation is difficult to predict, the prediction by the SP method may not be as accurate as the prediction without functional registration. However, if the shape of the curve to be predicted is of major concern, the SP method is a better approach.

4.4. Comparison with logarithm transformation methods

As has been discussed in the introduction, one feasible prediction approach for warping functions is to predict transformed warping functions. Such methods typically transform highly constrained warping functions to unconstrained functions, and then linear models are employed to predict the transformed functions. The transformations in such methods always incorporate "logarithm" and the original variation is shrunk or exaggerated. To show the superiority of the state-space model approach, it was compared with two transformation methods.

The first competitor method employs Jupp transformation. The method was considered in the warped regression model (Gernivi [8]). In this method, each

warping function $\gamma_n(t)$ is evaluated at fixed grids $0 = \gamma_{n0} < \gamma_{n1} < \cdots < \gamma_{nr} < \gamma_{n,r+1} = 1$, and each discretized vector is transformed by the Jupp transformation specified below:

$$\tau_{nj} = \log\left\{\frac{\gamma_{n,j+1} - \gamma_{n,j}}{\gamma_{n,j} - \gamma_{n,j-1}}\right\}, \qquad j = 1, \ldots, r,\ n = 1, \ldots, N.$$

VAR model is then employed to predict the transformed vectors. The last step is to transform the predicted vector $(\hat{\tau}_{N+1,1}, \ldots, \hat{\tau}_{N+1,r})$ back into a constrained increasing vector by the inverse Jupp transformation, say,

$$\begin{aligned}
\hat{\gamma}_{N+1,j} &= s_{N+1,j}/(1 + s_{N+1,r}),\\
s_{N+1,j} &= \sum_{k=1}^{j} \exp(\hat{\tau}_{N+1,1} + \cdots + \hat{\tau}_{N+1,k}), \qquad j = 1, \ldots, r.
\end{aligned}$$

This method typically requires fine grids so that the discretized vectors capture the major features of warping functions.

The second method is a functional approach. Similar to those transformations employed in Petersen and Müller [22], the transformation applied in this method has strict inverse only modulo the quotient space, and specifically, two functions $f_1(t)$ and $f_2(t)$ defined over $[0, 1]$ are equivalent if $f_1(t)/f_1(1) = f_2(t)/f_2(1)$. The transformation $\psi(\cdot)$ and its inverse are given as follows:

$$\begin{aligned}
r_n(t) \equiv \psi(\gamma_n)(t) &= \log\left(\dot{\gamma_n}\right)(t),\\
\psi^{-1}(r_n(t)) &= \int_0^t \frac{\exp\left(r_n(x)\right)}{\int_0^1 \exp(r_n(s))ds} dx.
\end{aligned}$$

The prediction method proposed by Aue et al. [2] was applied to predict the future transformed functions, and the predicted functions were then transformed back into warping functions with $\psi^{-1}(\cdot)$.

The warping functions were simulated under the second setup ($\beta = 0.8$), and $\{\gamma_i^e(t) \colon i \in \mathbb{N}\}$ were simulated in the same way (see Section 4.1.1) with 10 B-splines, say, $\gamma_i^e(t) = \sum_{j=1}^{10} \phi_{ij} B_j(t)$. The scores $\{\xi_{ij} \colon j = 1, \ldots, 9\}$ follow the following distribution

$$\begin{aligned}
&P(\xi_{ij} = 1) = P(\xi_{ij} = 2) = P(\xi_{ij} = 3) = P(\xi_{ij} = 4) = 1/4,\ j = 1, 2, 3,\\
&P(\xi_{ij} = 0) = q, P(\xi_{ij} = 1) = (1 - q)/2, P(\xi_{ij} = 2) = (1 - q)/2,\ j = 4, 5, 6,\\
&P(\xi_{ij} = 1) = P(\xi_{ij} = 2) = P(\xi_{ij} = 3) = P(\xi_{ij} = 4) = 1/4,\ j = 7, 8, 9.
\end{aligned}$$

Here, 500 warping functions were simulated. In the Jupp transformation method (JP), the warping functions were evaluated at 10 equally-spaced grids between 0 and 1. In the functional transformation approach (FT), 10 functional principal components were employed to represent the functions $\{r_n(t) \colon n = 1, \ldots, 500\}$. In our state space model method (MC), 10 prototypes were selected. The prediction error of $\hat{\gamma}(t)$ was evaluated with the spherical geodesic distance $d(\hat{\gamma}(t), \gamma(t)) =$

TABLE 4
Average prediction errors

Method \ q	0	0.3	0.5	1
MC	**0.055**	**0.095**	**0.117**	**0.045**
JP	0.078	0.122	0.155	0.289
FT	0.172	0.240	0.260	0.326

$\cos^{-1}(\langle S(\hat{\gamma}(t)), S(\gamma(t))\rangle)$. Table 4 displays the average prediction errors of different methods under different q's.

As q is large (close to 1), the middle part of the simulated warping functions is more likely to be flat, say, $\dot{\gamma}_n \approx 0$, and the "logarithm" transformations are more likely to exaggerate the original variation. Table 4 shows that the MC method is superior to the other two competitor methods, especially when the warping functions are flat over some intervals.

5. Analysis of pollution concentration trajectories

The SP method was applied to predict the air quality trajectories (De Vito et al. [23]). The data is available at the UCI Machine Learning Repository. The dataset contains hourly averaged observations collected from 5 metal oxide chemical sensors embedded in an Air Quality chemical multi-sensor device. The device was placed at road level in a significantly polluted area in an Italian city. The pollution concentration was recorded from March 2004 to February 2005 (one year). Here we analyzed the non metanic hydrocarbons concentration.

The pollution concentration is highly influenced by the traffic flow, so the trajectories share a common two-peak pattern (one peak in the morning and one peak in the afternoon). However, humidity, wind speed, temperature and other environmental factors can also influence the concentration, thus the trajectories display phase variation. Figure 5 displays the smoothed NMHC concentration trajectories and the registered trajectories. The trajectories of the weekdays are marked in black; those of Saturdays are marked in blue; and those of Sundays are marked in red. Figure 6 displays the (prototype) warping functions.

As the trajectories for the weekdays share a mean different from that of weekends, the amplitude functions were centralized with the means of each day of the week. After removing the days with too many missing values, there are 357 trajectories in total. We found that the SP method produced the overall best prediction when the amplitude and warping functions are predicted separately. The first 300 curves were used to train the model for predicting the rest 57 curves, and FAR(1) models were fitted to predict the amplitude functions of the trajectories to be predicted. The SP method was compared with the prediction method without functional registration (Aue et al. [2]).

Table 5 displays the average ℓ^2 prediction error and amplitude distance between the predicted curves and the corresponding smoothed curves under different numbers of states g and dimensions of eigen-space d. It is noted that the

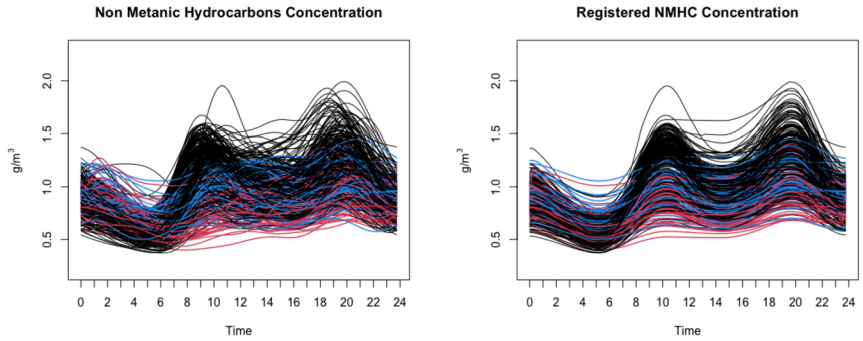

FIG 5. *Smoothed daily NMHC concentration and registered trajectories*

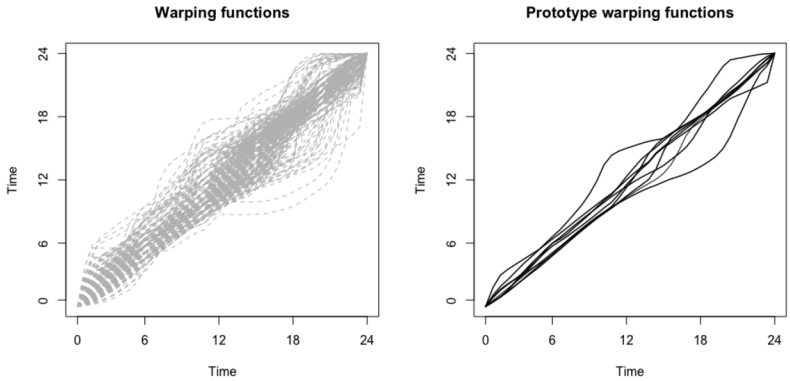

FIG 6. *Warping functions and prototype warping functions.*

TABLE 5
Prediction comparison. Format of each block: average ℓ^2 prediction error (average amplitude distance)

Shape-preserving method			
g \ d	3	5	7
6	141.805(0.856)	145.567(0.848)	147.846(0.856)
7	141.988(0.854)	145.507(0.846)	148.507(0.866)
8	142.608(0.851)	144.627(0.857)	141.010(0.857)
9	140.398(0.846)	144.824(0.854)	143.171(0.848)
10	**138.352(0.850)**	144.326(0.847)	145.268(0.847)
11	141.591(0.844)	146.282(0.847)	141.682(0.855)
Prediction without functional registration			
d	3	5	7
–	138.552(0.892)	136.458(0.879)	**134.998(0.880)**

SP method sacrifices marginal prediction accuracy to preserve the common two-peak pattern of the functional time series. The best SP prediction is achieved when $d = 3$, $g = 10$, while the competitor method reaches its best performance when $d = 7$. This is because the functional registration step assures that less functional principal components are needed to capture most of the vertical variation.

6. Conclusions

In this paper, we develop a new prediction method for stationary functional time series that display a common pattern. To the best of our knowledge, our SP method is the first to incorporate functional registration into prediction of functional time series. The prediction algorithm jointly predicts the amplitude and phase components. These two predicted components are then combined to form the final prediction.

The SP method has two main advantages. First, if the curves displayed a common pattern and significant phase variation, considering vertical variation only would lead to the loss of main features. Comparatively, the new methodology separates amplitude and phase components first, thus the SP method can preserve the pattern better. Second, the SP method is "natural" because: (1.) $S(\cdot)$ is a bijective transformation, thus no additional adjustments are needed to transform a SRSF back into a warping function, which avoids bias; (2.) The method does not directly apply linear models to non-linear objects, making the prediction natural and avoiding extremely small values resulting from the "logarithm". The simulation study and real data analysis of non metanic hydrocarbons concentration data show that the SP method is superior to the prediction methods without functional registration in capturing the common pattern of trajectories, and meanwhile produce predictions with competitive prediction accuracy. In this paper, it is assumed that the pattern (shape) repeats with a fixed period. However, in some cases, such as some biomedical or physical signals, a signal may be composed of multiple components, and each component repeats itself at different rate. The extension of the SP method to such cases is a research topic that will be pursued in the future.

Appendix A: Technical proofs

Proof of Proposition 3.1. By the Bayesian theorem and Assumption (A2), we deduce that

$$P(\hat{\omega}_{n+1}^{(f)}, \hat{\omega}_{n+1}^{(a)} | \hat{\omega}_n^{(f)}, \hat{\omega}_n^{(a)}) = \sum_{\omega_{n+1}^{(f)}, \omega_{n+1}^{(a)}, \omega_n^{(f)}, \omega_n^{(a)}} P(\hat{\omega}_{n+1}^{(f)}, \hat{\omega}_{n+1}^{(a)} | \omega_{n+1}^{(f)}, \omega_{n+1}^{(a)}) P(\omega_{n+1}^{(f)}, \omega_{n+1}^{(a)} | \omega_n^{(a)}, \omega_n^{(f)}) P(\omega_n^{(a)}, \omega_n^{(f)} | \hat{\omega}_n^{(a)}, \hat{\omega}_n^{(f)}),$$

and

$$P(\hat{\omega}_{n+1}^{(f)}, \hat{\omega}_{n+1}^{(a)} | \omega_{n+1}^{(f)}, \omega_{n+1}^{(a)}) = P(\hat{\omega}_{n+1}^{(a)} | \omega_{n+1}^{(a)}) P(\hat{\omega}_{n+1}^{(f)} | \omega_{n+1}^{(f)}),$$

$$P(\omega_n^{(a)},\omega_n^{(f)}|\hat{\omega}_n^{(a)},\hat{\omega}_n^{(f)}) = \frac{P(\hat{\omega}_n^{(a)},\hat{\omega}_n^{(f)}|\omega_n^{(a)},\omega_n^{(f)})P(\omega_n^{(a)},\omega_n^{(f)})}{P(\hat{\omega}_n^{(a)},\hat{\omega}_n^{(f)})}$$
$$= \frac{P(\hat{\omega}_n^{(a)}|\omega_n^{(a)})P(\hat{\omega}_n^{(f)}|\omega_n^{(f)})P(\omega_n^{(a)},\omega_n^{(f)})}{P(\hat{\omega}_n^{(a)},\hat{\omega}_n^{(f)})}$$
$$= \frac{P(\hat{\omega}_n^{(a)}|\omega_n^{(a)})P(\hat{\omega}_n^{(f)}|\omega_n^{(f)})P(\omega_n^{(a)},\omega_n^{(f)})}{\sum_{\omega_n^{(a)},\omega_n^{(f)}} P(\hat{\omega}_n^{(a)}|\omega_n^{(a)})P(\hat{\omega}_n^{(f)}|\omega_n^{(f)})P(\omega_n^{(f)},\omega_n^{(a)})}.$$

The result follows. □

The least squares estimator of P is defined as the minimizer of the quantity $\sum_{n=2}^{N}\|\hat{\omega}_n - \hat{\omega}_{n-1}P\|^2$, where $\hat{\omega}_n = \hat{\omega}_n^{(f)} \otimes \hat{\omega}_n^{(a)}$. By Proposition 3.1, we have $E(\hat{\omega}_{n+1}|\hat{\omega}_n) = \hat{\omega}_n\widetilde{P}$. Notationally, let

$$L_N(P) = \frac{1}{N}\sum_{n=2}^{N}\|\hat{\omega}_n - \hat{\omega}_{n-1}P\|^2, \qquad L(P) = E\{L_N(P)\} = E\{\|\hat{\omega}_2 - \hat{\omega}_1 P\|^2\}.$$

To prove Theorem 3.1, first we state the following lemma from van der Plas [29].

Lemma A.1. *Let $\{X_n : n \in \mathbb{N}\}$ be a stationary and ergodic process with values in a Euclidean space E, and Θ be a compact subspace of some Euclidean space. Let F be a real valued measurable function on $E \times \theta$ such that $F(x,\theta)$ is a continuous function of θ for all $x \in E$. Define $\phi(x) = \sup_{\theta\in\Theta}|F(x,\theta)|$ for all x and assume that $E\{\phi(X_0)\} < \infty$, then $\lim_{N\to\infty} N^{-1}\sum_{n=1}^{N} F(X_n,\theta) = E\{F(X_0,\theta)\}$, a.s. uniformly for all $\theta \in \Theta$.*

Proof of Theorem 3.1. The first part holds since P is in a compact set and L_N is a continuous function. We now show the second part. We deduce that

$$L(\widehat{P}_N) - L(\widetilde{P}) = E\{\|\hat{\omega}_2 - \hat{\omega}_1\widehat{P}_N\|^2\} - E\{\|\hat{\omega}_2 - \hat{\omega}_1\widetilde{P}\|^2\}$$
$$= E\{(\hat{\omega}_1\widehat{P}_N)^T\hat{\omega}_1\widehat{P}_N\} - E\{(\hat{\omega}_1\widetilde{P})^T\hat{\omega}_1\widetilde{P}\}$$
$$+ 2E\{\hat{\omega}_2^T(\hat{\omega}_1\widetilde{P} - \hat{\omega}_1\widehat{P}_N)\},$$

and

$$E\{\hat{\omega}_2^T(\hat{\omega}_1\widetilde{P} - \hat{\omega}_1\widehat{P}_N)\} = E[E\{\hat{\omega}_2^T(\hat{\omega}_1\widetilde{P} - \hat{\omega}_1\widehat{P}_N)|\hat{\omega}_1\}]$$
$$= E\{(\hat{\omega}_1\widetilde{P})^T(\hat{\omega}_1\widetilde{P} - \hat{\omega}_1\widehat{P}_N)\}.$$

Assume $\widehat{P}_N$ is obtained from an independent copy of samples, then if $\widehat{P}_N \neq \widetilde{P}$ and given $\widehat{P}_N$,

$$L(\widehat{P}_N) - L(\widetilde{P}) = E\{(\hat{\omega}_1\widehat{P}_N)^T\hat{\omega}_1\widehat{P}\} - E\{(\hat{\omega}_1\widetilde{P})^T\hat{\omega}_1\widetilde{P}\}$$
$$+ 2E\{(\hat{\omega}_1\widetilde{P})^T(\hat{\omega}_1\widetilde{P} - \hat{\omega}_1\widehat{P}_N)\}$$
$$= \widetilde{P}^T\Sigma_\omega\widetilde{P} + \widehat{P}^T\Sigma_\omega\widehat{P} - 2\widetilde{P}^T\Sigma_\omega\widehat{P}_N$$

$$= \|(\Sigma_\omega)^{\frac{1}{2}}\widetilde{P} - (\Sigma_\omega)^{\frac{1}{2}}\widehat{P}_N\|^2 > 0,$$

where Σ_ω is the covariance matrix of $\hat{\omega}_n$. Since $\{\omega_n \colon n \in \mathbb{N}\}$ is ergodic and stationary, $\{\hat{\omega}_n \colon n \in \mathbb{N}\}$ is also ergodic and stationary. Hence by Lemma A.1,

$$\begin{aligned} 0 < L(\widehat{P}_N) - L(\widetilde{P}) &= L(\widehat{P}_N) - L_N(\widehat{P}_N) + L_N(\widehat{P}_N) - L(\widetilde{P}) \\ &\le L(\widehat{P}_N) - L_N(\widehat{P}_N) + L_N(\widetilde{P}) - L(\widetilde{P}) \\ &\le 2\sup_P |L(P) - L_N(P)| \to 0, \qquad a.s.. \end{aligned}$$

Therefore, $L(\widehat{P}_N) \to L(\widetilde{P})$ $a.s.$. Since $L(P)$ is a continuous function, $\widehat{P}_N \to \widetilde{P}$ a.s.. □

Before proving Theorem 3.2, we first define $F_n = \langle \hat{\omega}_n - \frac{1}{2}\hat{\omega}_{n-1}P, \hat{\omega}_{n-1}P\rangle$, and evidently, we have the relationship

$$F_i(n,\theta) = \frac{\partial F_n}{\partial \theta_i}, \qquad \frac{\partial L_N(P)}{\partial \theta_i} = -2N^{-1}\sum_{n=1}^{N} F_i(n,\theta).$$

In addition, we need to establish the following lemmas.

Lemma A.2. *Under Assumption (A1)–(A3), assume there exists a probability distribution π such that $\|P(\omega_n \in \cdot) - \pi(\cdot)\| \to 0$, as $n \to \infty$. Then $\{F_n\}$ is strong mixing for any initial distribution of F_0, and the mixing coefficients satisfy $\sum_{m=1}^{\infty} \alpha(m) < \infty$.*

Proof. Define $\mathcal{F}_m^n = \sigma(\omega_m, \ldots, \omega_n)$, $\widehat{\mathcal{F}}_m^n = \sigma(F_m, \ldots, F_n)$, and assume $\widehat{E} \in \widehat{\mathcal{F}}_0^n$ and $\widehat{F} \in \widehat{\mathcal{F}}_{n+m}^{\infty}$, then by Assumption (A2), we have

$$\begin{aligned} P(\widehat{E} \cap \widehat{F}) &= E(\mathbb{1}_{\widehat{E}}\mathbb{1}_{\widehat{F}}) = E\{E(\mathbb{1}_{\widehat{E}}\mathbb{1}_{\widehat{F}}|\mathcal{F}_0^{n+m-1})\} \\ &= E\{E(\mathbb{1}_{\widehat{E}}|\mathcal{F}_0^n)E)\mathbb{1}_{\widehat{F}}|\mathcal{F}_0^{n+m-1})\} \\ &= E\{g(E)h(\omega_{n+m-1})\} = E[E\{g(E)h(\omega_{n+m-1})|\sigma(\omega_n)\}] \\ &= E[E\{g(E)|\sigma(\omega_n)\}E\{h(\omega_{n+m-1})|\sigma(\omega_n)\}] \\ &= E\{\widetilde{g}(\omega_n)P^{m-1}h(\omega_n)\}, \end{aligned}$$

where

$$\begin{aligned} &g(E) = E(\mathbb{1}_{\widehat{E}}|\mathcal{F}_0^n),\ h(\omega_{n+m-1}) = E(\mathbb{1}_{\widehat{F}}|\mathcal{F}_0^{n+m-1}), \\ &\widetilde{g}(\omega_n) = E\{g(E)|\sigma(\omega_n)\},\ P^m h(x) = E\{h(\omega_m)|\omega_0 = x\}. \end{aligned}$$

By similar arguments, $P(\widehat{E}) = E\{\widetilde{g}(\omega_n)\}$, $P(\widehat{F}) = E\{P^{m-1}h(\omega_n)\}$. Therefore we have,

$$\begin{aligned} P(\widehat{E} \cap \widehat{F}) - P(\widehat{E})P(\widehat{F}) &= E\{\widetilde{g}(\omega_n)P^{m-1}h(\omega_n)\} - E\{\widetilde{g}(\omega_n)\}E\{P^{m-1}h(\omega_n)\} \\ &= E[\widetilde{g}(\omega_n)\{P^{m-1}h(\omega_n) - \pi(h)\}] \end{aligned}$$

$$+ E\{\tilde{g}(\omega_n)\}[\pi(h) - E\{P^{m-1}h(\omega_n)\}],$$

where $\pi(h) = \int h(x)\pi(dx)$. Since $E\{\tilde{g}(\omega_n)\}$ is bounded by 1, we have as $m \to \infty$,

$$\begin{aligned}\alpha(m) &= \sup_{\widehat{E},\widehat{F}} |P(\widehat{E} \cap \widehat{F}) - P(\widehat{E})P(\widehat{F})| \\ &\le 2 \sup_{\omega} E[\|P^{m-1}(h(\omega_n) \in \cdot) - \pi(h(\cdot))\|] \to 0.\end{aligned}$$

Because $P^{m-1}(\omega_n \in \cdot)$ converges to $\pi(\cdot)$ exponentially fast by the property of Markov chain, we conclude $\sum_{m=1}^{\infty} \alpha(m) < \infty$. □

We now show that $\sqrt{N}(\widehat{\theta}_N - \tilde{\theta})$ converges to normality as $N \to \infty$, based on the following results (Ibraginov [15]), which establishes the asymptotic normality for bounded univariate strong mixing processes.

Lemma A.3. *Suppose that the stationary process $\{F_n\}$ is a strong mixing sequence with mixing coefficient $\alpha(m)$. If the random variable ξ is measurable with respect to $\sigma(F_{-\infty}, \ldots, F_n)$, and the random variable η is measurable with respect to $\sigma(F_{n+m}, \ldots, F_{\infty})$, and if $|\xi| < C_1$, $|\eta| < C_2$, then $\operatorname{cov}(\xi, \eta) \le 4C_1C_2\alpha(m)$.*

Lemma A.4. *Let $\{X_n : n \in \mathbb{N}\}$ be a centered strictly stationary, strong mixing sequence. Suppose that there exists $B < \infty$ such that $|X_n| < B$ a.s. and $\sum_{m=1}^{\infty} \alpha(m) < \infty$. Then $\sigma^2 = E(X_0^2) + 2\sum_{j=1}^{\infty} E(X_0X_j) < \infty$ and if $\sigma > 0$, as $N \to \infty$, $N^{-1/2}S_N \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$, where $S_N = \sum_{n=1}^{N} X_n$.*

Proof of Theorem 3.2. Note that $F_i(n, \theta)$ is a centered stochastic process:

$$\begin{aligned}E\{F_i(n, \tilde{\theta})\} &= E\left\{\left\langle \hat{\omega}_n - \hat{\omega}_{n-1}\widetilde{P}, \frac{\partial \hat{\omega}_{n-1}P}{\partial \theta_i}\bigg|_{\widetilde{P}} \right\rangle\right\} \\ &= E\left\{E\left\{\left\langle \hat{\omega}_n - \hat{\omega}_{n-1}\widetilde{P}, \frac{\partial \hat{\omega}_{n-1}P}{\partial \theta_i}\bigg|_{\widetilde{P}} \right\rangle\right\} \bigg| \hat{\omega}_{n-1}\right\} \\ &= E\left\{\left\langle E\left\{\hat{\omega}_n - \hat{\omega}_{n-1}\widetilde{P} \bigg| \hat{\omega}_{n-1}\right\}, \frac{\partial \hat{\omega}_{n-1}P}{\partial \theta_i}\bigg|_{\widetilde{P}} \right\rangle\right\} = 0.\end{aligned}$$

Therefore, by the result in Lemma A.2, $\{F_i(n, \tilde{\theta}) : n \in \mathbb{N}\}$ is a centered, strong mixing, bounded and stationary sequence. Then by Lemma A.4, we have

$$\sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N} F_i(n, \tilde{\theta})\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_{ii}^2),$$

where $\sigma_{ii} = E\{F_i^2(0, \tilde{\theta})\} + 2\sum_{k=1}^{\infty} E\{F_i(0, \tilde{\theta})F_i(k, \tilde{\theta})\}$, $i = 1, \ldots, g^2$. For the asymptotic normality of $N^{-1/2}\sum_{n=1}^{N} \frac{\partial F_n(\theta)}{\partial \theta}$, we only need to show that the covariance elements are finite. For $i \ne j$, define

$$\sigma_{N,ij} = E\left\{\sum_{n=1}^{N} F_i(n, \theta) \times \sum_{n=1}^{N} F_j(n, \theta)\right\}, \qquad R_{ij}(m) = E\{F_i(0, \theta)F_j(m, \theta)\},$$

since $F_i(n,\theta) \in \sigma(F_{-\infty},\dots,F_n)$, $F_j(n+m,\theta) \in \sigma(F_{n+m-1},\dots,F_\infty)$, then by Lemma A.3, we have

$$\sigma_{N,ij} = N\left(R_{ij}(0) + 2\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)R_{ij}(k)\right)$$
$$\leq N\left(R_{ij}(0) + \text{const.}\sum_{k\geq 1}\alpha(k)\right),$$

and thus

$$\text{cov}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}F_i(n,\theta), \frac{1}{\sqrt{N}}\sum_{n=1}^{N}F_j(n,\theta)\right) < \infty.$$

Observe that

$$\frac{\partial L_N(P)}{\partial\theta} = -\frac{2}{N}\sum_{n=1}^{N}\frac{\partial F_n(\theta)}{\partial\theta},$$

so by Lemma A.4 and the previous arguments, we have

$$N^{1/2}\left\{\frac{\partial L_N(P)}{\partial\theta}\bigg|_{\widetilde{P}}\right\} \xrightarrow{\mathcal{L}} \mathcal{N}(0,4\Sigma). \tag{B1}$$

By the mean value theorem, we deduce that

$$0 = N^{1/2}\left\{\frac{\partial L_N(P)}{\partial\theta}\bigg|_{\widetilde{P}}\right\} + \left\{\frac{\partial^2 L_N(P)}{\partial\theta\partial\theta'}\bigg|_{P_N^*}\right\}N^{1/2}(\hat{\theta}_N - \tilde{\theta}), \tag{B2}$$

where P_N^* is a stochastic $g\times g$ transition probability matrix satisfying $\|P_N^* - \widetilde{P}\| \leq \|\widehat{P}_N - \widetilde{P}\|$. By Lemma A.1, $\frac{\partial^2 L_N(P)}{\partial\theta\partial\theta'}$ converges uniformly for all P. Evidently,

$$E\left\{\frac{\partial^2 L_N(P)}{\partial\theta_i\partial\theta_j}\bigg|_{\widetilde{P}}\right\} = 2a_{ij},$$

thus we have

$$\left\|\frac{\partial^2 L_N(P)}{\partial\theta\partial\theta'}\bigg|_{\widehat{P}_N} - 2A\right\|^2 \leq 2\left\{\left\|\frac{\partial^2 L_N(P)}{\partial\theta\partial\theta'}\bigg|_{\widehat{P}_N} - \frac{\partial^2 E(L_N)(P)}{\partial\theta\partial\theta'}\bigg|_{\widehat{P}_N}\right\|^2 + \left\|\frac{\partial^2 E(L_N)(P)}{\partial\theta\partial\theta'}\bigg|_{\widehat{P}_N} - \frac{\partial^2 E(L_N)(P)}{\partial\theta\partial\theta'}\bigg|_{\widetilde{P}}\right\|^2\right\}.$$

The first summand converges to zero almost surely by Lemma A.1, and the second summand converges to zero by Theorem 3.1 and the fact that L_N is a continuous function. Consequently,

$$\frac{\partial^2 L_N(P)}{\partial\theta\partial\theta'}\bigg|_{P_N^*} \xrightarrow{a.s.} 2A \qquad \text{a.s..}$$

Then the theorem follows immediately from (B1) and (B2). □

Acknowledgments

We are grateful to the Associate Editor and two referees for their comments and suggestions that led to substantial improvement of the paper.

References

[1] Antoniadis, A. and Sapatinas, T. (2003). Wavelet methods for continuous time prediction using Hilbert-valued autoregressive processes. *Journal of Multivariate Analysis* **87** 113–158. MR2007265

[2] Aue, A., Norinho, D. D. and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association* **110** 378–392. MR3338510

[3] Besse, P. C., Cardot, H. and Stephenson, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics* **27** 673–687.

[4] Brumback, L. C. and Lindstrom, M. J. (2004). Self modeling with flexible, random time transformations. *Biometrics* **60** 461–470. MR2066281

[5] de Boor, C. (1978). A practical guide to splines. Springer-Verlag, New York-Berlin. MR0507062

[6] Cheng, M. Y. and Wu, H. T. (2013). Local linear regression on manifolds and its geometric interpretation. *Journal of the American Statistical Association* **108** 1421–1434. MR3174718

[7] Dai, X. and Müller, H. G. (2018). Principal component analysis for functional Data on Riemannian manifolds and spheres. *Annals of Statistics* **46** 3334–3361. MR3852654

[8] Gervini, D. (2015). Warped functional regression. *Biometrika* **102** 1–14. MR3335092

[9] Hornik, K., Feinerer, I., Kober, M. and Buchta, C. (2012). Spherical K-means clustering. *Journal of Statistical Software* **50** 1–22.

[10] Horváth, L. and Kokoszka, P. (2012). Inference for functional data with applications. Springer Series in Statistics. Springer, New York. MR2920735

[11] Hyndman, R. J. and Shang, H, L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society* **38** 199–211. MR2750314

[12] Lin, Y. T., Malik, J. and Wu, H. T. (2019). Wave-shape oscillatory model for biomedical time series with applications. *arXiv preprint* arXiv:1907.00502.

[13] Lin, C. Y., Su, L. and Wu, H, T. (2018). Wave-shape function analysis. *Journal of Fourier Analysis and Applications* **24** 451–505. MR3776331

[14] Lütkepohl, H. (2005). New introduction to multiple time series analysis. Springer-Verlag, Berlin. MR2172368

[15] Ibraginov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability and its Applications* **7** 349–382. MR0148125

[16] Jiao, S., Aue, A. and Ombao, Hernando. (2021). Functional time series prediction under partial observation of the future curve. *Journal of the American Statistical Association* DOI: 10.1080/01621459.2021.1929248.

[17] Jupp, David L. B. (1978). Approximation to data by splines with free knots. *SIAM Journal of Numerical Analysis* **15** 328–343. MR0488884
[18] Kargin, V. and Onatski, A. (2008). Curve forecasting by functional autoregression. *Journal of Multivariate Analysis* **99** 2508–2526. MR2463404
[19] Kneip, A. and Ramsay, J. O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association* **103** 1155–1165. MR2528838
[20] Kokoszka, P., Miao, H., Petersen, A. and Shang, H. L. (2019). Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting* **35** 1304–1317.
[21] Krafty, R. T., Gimotty, P. A., Holtz, D., Coukos, G. and Guo, W. (2008). Varying coefficient model with unknown within-subject covariance for analysis of tumor growth curves. *Biometrics* **64** 1023–1031. MR2522249
[22] Perterson, A. and Müller, H. G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics* **44** 183–218. MR3449766
[23] De Vito, S., Massera, E., Piga, M, Martinotto, L. and Di Francia, G. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical* **129** 750–757.
[24] Sentürk, D. and Müller, H. G. (2008). Generalized varying coefficient models for longitudinal data. *Biometrika* **95** 653–666. MR2443181
[25] Sentürk, D. and Müller, H. G. (2010). Functional varying coefficient models for longitudinal data. *Journal of the American Statistical Association* **105** 1256–1264. MR2752619
[26] Srivastava, A. and Klassen, E. P. (2016). Functional and shape data analysis. Springer-Verlag, New York. MR3821566
[27] Su, J., Kurtek, S., Klassen, E. P. and Srivastava, A. (2014). Statistical analysis of trajectories on Riemannian manifolds: Bird migration, hurricane tracking and video surveillance. *The Annals of Applied Statistics* **8** 530–552. MR3192001
[28] Ramsay, J. O. and Silverman, B. W. (2005). Functional data analysis. Springer Series in Statistics. Springer, New York. MR2168993
[29] Van Der Plas A. P. (1983). On the estimation of the parameters of Markov probability models using macro data. *Annals of Statistics* **11** 78–85. MR0684865